DISSERTATIONES LINGUISTICAE UNIVERSITATIS TARTUENSIS

8

DISSERTATIONES LINGUISTICAE UNIVERSITATIS TARTUENSIS
8

KÕNE AJALISE STRUKTUURI MODELLEERIMINE EESTIKEELSELE TEKST-KÕNE SÜNTEESILE

MODELLING THE TEMPORAL STUCTURE OF SPEECH FOR THE ESTONIAN TEXT-TO-SPEECH SYNTHESIS

MEELIS MIHKLA

TARTU ÜLIKOOLI
KIRJASTUS

Eesti Keele Instituut,
Doktorikool „Keeleteadus ja -tehnoloogia”,
Tartu Ülikool, Eesti ja Üldkeeleteaduse Instituut

Väitekiri on kaitsmisele suunatud Tartu Ülikooli Eesti ja Üldkeeleteaduse Instituudi nõukogu otsusega 04.12.2007.

Juhendajad: Einar Meister, filosoofiadoktor, TTÜ Küberneetika Instituut
Haldur Õim, filoloogiadoktor, Tartu Ülikool

Oponendid: Toomas Altosaar, tehnikateaduste doktor, Helsingi Tehnikaülikool
Diana Krull, filoloogiadoktor, Stockholmi Ülikool

Kaitsmine toimub 15. jaanuaril 2008 kell 14.15 Tartu Ülikooli nõukogu saalis.

Töö valmimist on toetanud ja trükikulud on katnud keeleteaduse ja -tehnoloogia doktorikool ning Eesti Keele Instituut.

ISSN 1024–395X
ISBN 978–9949–11–797–0 (trükis)
ISBN 978–9949–11–798–7 (PDF)

Tartu Ülikooli Kirjastus
www.tyk.ee
Tellimus nr 546

LÜHIKOKKUVÕTE

Käesolevas väitekirjas esitatakse metodoloogia eestikeelse kõne ajalise struktuuri mudelite automaatseks genereerimiseks kõrgekvaliteedilisele tekst-kõne sünteesile. Kõne prosoodia modelleerimise põhilised probleemid on olnud seotud „ähmase piirkonnaga" kõne diskreetse sümbolesituse ja pideva kõnelaine vahel. Pole ju tavalises kirjalikus tekstis peale kirjavahemärkide kõne ajalise struktuuri kohta (kõneüksuste ja pauside kestused, pauside asukohad, kõnetempo jms) ühtegi suunavat märki. Ajalise struktuuri loomulikkus kõnesünteesis eeldab, et me oskame edasi anda häälikute ja pauside kestusi ning pauside paiknemist kõnevoos nii, et nende väärtused oluliselt ei erine väärtustest sidusas kõnes. Reeglipõhiste prosoodiamudelite puuduseks tekst-kõne sünteesis on olnud asjaolu, et reeglid põhinevad paljuski nn „laboratoorse kõne" mõõtmiste alusel tehtud üldistustel ja et neis ilmneb vigu sõltumatult tuletatud reeglite samaaegsest rakendamisest. Sidusa kõne korpuste kasutamine ja statistiline optimeerimine võimaldavad reeglite kirjutamise asendada statistilise modelleerimisega ja seega parandada sünteeskõne kvaliteeti.

Sidusa kõne korpustele rakendati töös erinevaid statistilisi meetodeid (lineaarset ja logistilist regressiooni, klassifikatsiooni ja regressioonipuid (CART) ning närvivõrke) häälikute ja pauside kestuste prognoosimiseks. Kuna eesmärgiks on tekst-kõne sünteesile kõne ajalise struktuuri modelleerimine, siis moodustasid sidusa kõne korpuse erinevat tüüpi etteloetud tekstide (ilukirjandus, uudised, tekstid eesti keele foneetilisest andmebaasist) salvestused 27 diktori esituses.

Modelleerimiseksperimentidel leiti, et pauside kestused ja nende asukohad kõnevoos on prognoositavad. Mudelid osutusid kõige tugevamalt seotuks teksti liigendusega (kirjavahemärkide ja sidesõnadega), aga ka kaugusega eelmisest pausist ja asendist lauses. Etteloetud tekstides on pausid kestuse poolest klassifitseeritavad, nad on automaatselt liigitatavad lõigu-, lause- ja fraasilõpu pausideks.

Segmentaalkestuste prognoosimisel osutusid olulisteks tunnused, mis kirjeldavad vaadeldava foneemi mõjutust naaberfoneemidest, aga samuti tunnused, mis iseloomustavad foneemi paiknemist lausungi hierarhilises struktuuris (nt foneemi asend silbis, sõna asend fraasis jms). Lisaks on statistiliselt tähtsad needki tunnused, mis iseloomustavad foneemi klassi, silbi rõhulisust, sõna ühesilbilisust ja teksti süntaktilist liigendust.

Eesti keeles on sõnal ja tema vormil tähtis roll nii grammatikas kui ka foneetikas. Töös tuvastati, et sõna moodustavate segmentide kestusi mõjutavad sõnade süntaktilised, morfoloogilised ja sõnaliigi tunnused.

Erinevate prognoosimeetodite võrdlemisel ilmnes, et lineaarne regressioon on prognoositäpsuselt kestuste ennustamisel võrdväärne statistiline meetod mittelineaarsete meetoditega (CART'i ning närvivõrkudega).

Kõne ajalise struktuuri korpuspõhine modelleerimine pakub lisaks kõnetehnoloogiale huvi ka foneetikale, sest meetod võimaldab näiteks analüüsida väikesi, varjatud, kuid olulisi erinevusi häälikute kestustes, mis tulenevad sõna morfoloogilis-süntaktilisest liigendusest ja sõnaliigist. Korpuspõhine statistiline metodoloogia võimaldab foneetikateadustes testida suurtel andmehulkadel erinevaid teoreetilisi lähenemisi ja teha paljude nähtuste täppisanalüüsi, mis annab statistiliselt põhjendatud aluse tunnetuslike mehhanismide toimimisest foneetikas.

TÄNU

Väitekiri on valminud aastatel 2004–2007 Eesti Keele Instituudis, aastast 2005 ka doktorikooli „Keeleteadus ja -tehnoloogia" raames. Töö valmimisele on mõistuse ja südamega kaasa aidanud suur hulk inimesi.

Esimesena tänan oma juhendajaid dr Einar Meistrit ja prof Haldur Õimu väärtuslike nõuannete eest nii teemaarenduses kui ka doktorikooli õpingutes. Autorit seob Einar Meistriga lisaks doktoritööle ka pikaajaline viljakas koostöö eestikeelse tekst-kõne sünteesi vallas. Eriline tänu dr Arvo Eegile väärtuslike märkuste eest nii mitmete artiklite kui ka väitekirja kokkuvõtte kirjutamisel. Olulist abi sain Arvo Eegilt just väitekirjas kasutatud mõistete ja kontseptsioonide formuleerimisel ja kirjeldamisel.

Tänan ka artiklite kaasautoreid Hille Pajupuud ja Krista Kerget lausete intonatsiooni uuringutes ning Jüri Kuusikut, kelle nõu ja abiga ma statistiliste prognoosimeetoditeni jõudsin.

Eesti Keele Instituudi direktor prof Urmas Sutrop innustas mind nii doktorikooli õpingutes kui ka töö jaoks olulise artikli [P8] kirjutamisel ajakirjale „Trames". Doktorikooli juhataja prof Karl Pajusalu andis väga vajalikke näpunäiteid artikliväitekirja kokkuvõtte kirjutamiseks.

Autorit seob ka tihe koostöö Põhja-Eesti Pimedate Ühinguga. Pimedad ja vaegnägijad on eestikeelse tekst-kõne sünteesi igapäevased kasutajad arvutikeskkonnas. Nemad on ka kõneprosoodia modelleerimistulemuste parimad testijad. Artur Räpilt ja Eduard Borissenkolt olen saanud kõne ajalise struktuuri mudelite toimimise kohta tagasisidet märkuste ja soovitustena. Aitäh teile!

Tänan Sirje Ainsaart artiklite kvaliteetsete inglisekeelsete tõlgete, Jana Tiitust Tallinna Ülikoolist kokkuvõtte kiire ja asjatundliku tõlkimise ning Eva-Liina Asu-Garcia Tartu Ülikoolist inglisekeelse osa korrektuuri eest. Tänuga tahaksin nimetada kolleege Liisi Piitsa ja Indrek Kiisselit, kes oma ametikohustuste kõrvalt jõudsid tööd kriitiliselt lugeda ja trükiks ette valmistada.

Eriline tänu ka minu perele: abikaasa Külli ja tütred Triin, Maarja, Laura ja Liisa on kõik minu hiliseid õpinguid mõistvalt toetanud.

Suur tänu kõigile kolleegidele ja kaastöölistele, kes on kaasa aidanud selle töö valmimisele ja vormistamisele.

Tallinn, detsember 2007
Meelis Mihkla

SISUKORD

PUBLIKATSIOONIDE NIMEKIRI

Allpool esitatud publikatsioonid on väitekirja aluseks ja neile on tekstis viidatud vastava numberloendiga [P1]…[P8].

[P1] Mihkla, Meelis; Pajupuu, Hille; Kerge, Krista; Kuusik, Jüri 2004. Prosody modelling for Estonian text-to-speech synthesis. – The First Baltic Conference. Human Language Technologies, The Baltic Perspective, April 21–22 2004. Riga: 127–131.

[P2] Mihkla, Meelis; Kuusik, Jüri 2005. Analysis and modelling of temporal characteristics of speech for Estonian text-to-speech synthesis. Linguistica Uralica, XLI(2): 91–97.

[P3] Mihkla, Meelis 2005. Modelling pauses and boundary lengthenings in synthetic speech. – Proceedings of the Second Baltic Conference on Human Language Technologies, April 4–5, 2005. Tallinn: 305–310.

[P4] Mihkla, Meelis; Kerge, Krista; Pajupuu, Hille 2005. Statistical modelling of intonation and breaks for Estonian text-to-speech synthesizer. – Proceedings of the 16th Conference of Electronic Speech Signal Processing, joined with the 15th Czech-German Workshop "Speech Processing", Robert Vich (Toim.), September 26–28. Prague: 91–98, Dresden: TUDpress.

[P5] Mihkla, Meelis 2006. Pausid kõnes. Keel ja Kirjandus, XLIX(4): 286–295.

[P6] Mihkla, Meelis 2006. Comparision of statistical methods used to predict segmental durations. – The Phonetics Symposium 2006: Fonetiikan Päivät 2006, Helsingi, 30.–31.08.2006. (Toim.) Aulanko, Reijo; Wahlberg, Leena; Vainio, Martti. Helsingi: 120–124, University of Helsinki.

[P7] Mihkla, Meelis 2007. Morphological and syntactic factors in predicting segmental durations for Estonian text-to-speech synthesis. – Proceedings of the 16th International Congress of Phonetic Sciences, Saarbrücken, 6–10 August 2007, (Toim.) Jürgen Trouvain, William J. Barry. Saarbrücken: 2209–2212.

[P8] Mihkla, Meelis 2007. Modelling speech temporal structure for Estonian text-to-speech synthesis: feature selection. Trames. Journal of the Humanities and Social Sciences, 11(3): 284–298.

1. SISSEJUHATUS

1.1. Töö eesmärgid

Üheks oluliseks märksõnaks kõnetehnoloogias on kõne variatiivsus. Kui kõnetuvastuses põhjustab kõnelaine variatiivsus sageli probleeme, siis kõnesünteesis viib vähene variatiivsus sünteeskõne monotoonsusele ja ebaloomulikkusele (Tatham, Morton 2005:9). Kõne ajalise struktuuri loomulikkus eeldab seda, et me oskame häälikute ja pauside kestuste variatiivsust ning pauside paiknemist kõnevoos võimalikult hästi sünteeskõnes edasi anda.

Käesoleva uurimuse põhiliseks motivatsiooniks oli aastatel 1997–2002 välja töötatud eestikeelse tekst-kõne süntesaatori väljundkõne suhteline monotoonsus ja halb sidusus. See süntesaator põhines reeglipõhisele prosoodiamudelile (Mihkla, Meister, Eek 2000). Reeglipõhiste mudelite puudus on, et nad põhinevad paljuski nn „laboratoorse kõne" mõõtmiste baasil tehtud üldistustel ja neis ilmneb vigu sõltumatult tuletatud reeglite samaaegsel rakendamisel. Suurte kõnekorpuste kasutamine ja statistiline optimeerimine võimaldab aga reeglite kirjutamise asendada statistilise modelleerimisega ja parandada sünteeskõne kvaliteeti (Sagisaka 2003).

Töö eesmärgiks on töötada välja metodoloogia kõne ajalise struktuuri mudelite automaatseks genereerimiseks kõrgekvaliteedilisele tekst-kõne sünteesile. Selleks rakendati sidusa kõne korpustele erinevaid statistilisi meetodeid (lineaarset ja logistilist regressiooni, CART meetodit ja närvivõrke) kõneüksuste (so häälikute ja pauside) kestuste prognoosimiseks. Neid statistilisi tehnikaid on plaanis rakendada kõneprosoodia genereerimisel eestikeelsete korpuspõhiste süntesaatorite jaoks, mis põhinevad muutuva pikkusega kõneüksuste valikualgoritmidel (Mihkla jt 2007). Kõne ajalise struktuuri korpuspõhine modelleerimine pakub huvi ka foneetikas, sest ta võimaldab analüüsida väikesi, varjatud, kuid olulisi erinevusi häälikukestustes, mis tulenevad sõnaliigist [P7]. Arvatakse, et korpuspõhine statistiline lähenemine saab enamlevinuks foneetikateadustes, sest ta võimaldab erinevaid teoreetilisi lähenemisi testida suurtel andmehulkadel ja teha täppisanalüüsi, mis annab statistiliselt põhjendatud aluse tunnetuslike regulatsioonimehhanismide toimimisest foneetikas.

1.2. Töö ülesehitus

Väitekiri koosneb tutvustavast osast ja 8 artikli koopiast. Tutvustav osa on jagatud seitsmesse peatükki.

I peatükis on käesolev sissejuhatus, kus tutvustatakse töö problemaatikat ja ülesehitust, esitatakse publikatsioonide lühiülevaade koos autori panuse selgitamisega kaasautorlusega artiklites ning tutvustatakse mõisteid ja kontseptsioone, mis on seotud kõne ajalise struktuuri esitusega.

II peatükis antakse ülevaade kõnesünteesi strateegiatest, kõne ajalise regulatsiooni teooriatest ning faktorite ja tunnuste valiku põhialustest kõne ajastuse modelleerimisel.

III peatükis on lühiülevaade eestikeelse kõne ajalise struktuuri uurimustest: väldete käsitlemisest, häälikute mikroprosoodilistest tunnustest (omakestustest) ja pauside ning pausieelsete pikenduste uurimistöödest.

IV peatükis kirjeldatakse töödes kasutatud andmeid.

V peatükk on pühendatud statistilistele meetoditele, mida kasutati kestuste prognoosimisel. Samuti antakse ülevaade töödes kasutatud statistikaprogrammipakettidest.

VI peatükis kirjeldatakse arvukatel modelleerimiseksperimentidel saadud tulemusi, sealhulgas pauside kestusi ja pauside asukoha prognoosimist kõnevoos. Selekteeritakse olulisi tunnuseid segmentaalkestuste modelleerimiseks ja analüüsitakse sellega seotud sõnaprosoodia küsimusi. Kirjeldatakse erinevaid statistilisi mudeleid ning testitakse mudelite olulisust ja prognoositäpsust. Esitatakse meetodite võrdlus segmentaalsete kestuste modelleerimisel.

Kokkuvõte ja edasised töösuunad on toodud peatükis VII.

1.3. Artiklite lühiülevaade ja autori panusest kaasautorlusega töödes

Väitekiri põhineb 8 teaduslikule artiklile. Järgnevas on toodud artiklite lühiülevaade ja kirjeldus autori panusest kaasautorlusega töödes. [P1], [P2] ja [P4] kaasautoritele on tutvustatud neid kirjeldusi ning nad on nende sisuga nõus olnud.

[P1]-s käsitletakse eesti keele tekst-kõne süntesaatori prosoodia modelleerimise küsimusi: *kas*-küsimuse intonatsiooni modelleerimist, esimesi tähelepanekuid pauside ja pausieelsete pikenduste seostest teksti liigendusega ja esimest häälikute kestuste modelleerimist regressioonanalüüsi kasutades. Autori kirjutatud on pause ja pausieelseid pikendusi analüüsiv osa, samuti valmistas ta modelleerimisandmed ette ja interpreteeris tulemusi.

[P2]-s tutvustatakse kõnesünteesi jaoks segmentaalkestuste statistilist modelleerimist, kasutades seejuures regressioonanalüüsi. Autorilt pärineb pauside

analüüs ja pauside seos teksti liigendusega. Autor osales ka regressioonmudeli jaoks materjali ettevalmistamisel ja oluliste tunnuste kohta ekspertarvamuste kogumisel ning nende esitamisele regressioonanalüüsi kontekstis.

[P3]-s keskendutakse pauside ja pausieelsete pikenduste analüüsile sidusas kõnes ja pauside ning nende asukoha modelleerimisele kõnevoos. Autor oli artikli kirjutajaks ja eksperimentide läbiviijaks. Jüri Kuusik konsulteeris logistilise regressiooni rakendamist sisendandmetele.

[P4]-s modelleeritakse lineaarse regressiooni meetodit kasutades intonatsiooni morfoloogiliste, süntaktiliste ja sõnaliigi tunnuste alusel ning analüüsitakse pause ja kõnehingamist. Pause käsitletakse prosoodilise rühma piire markeerivate üksustena. Autor keskendus teooriale ja põhitooni statistilisele modelleerimisele ning sellega seotud kõnematerjali analüüsile. H. Pajupuu analüüsis pause ja hingamist kõnevoos ja määras lauserõhke. K. Kerge tegi lausete süntaktilist analüüsi ja interpreteeris saadud mudeleid.

[P5]-s ainuautorlusega artiklis on esitatud pikem käsitlus pausidest eesti-keelses kõnes ja pauside kestuse modelleerimisest klassikalise regressioonanalüüsi, klassifikatsiooni ja regressioonipuu (CART) meetodi ja närvivõrkude alusel. Pauside asukoha prognoosimine toimus logistilise regressiooni abil.

[P6]-s on autor võrrelnud erinevaid statistilisi prognoosimeetodeid (lineaarset regressiooni, CART-meetodit ja närvivõrke) prognoosivea, mudeli interpreteeritavuse, andmete eeltöötluse, jm kriteeriumide seisukohast.

[P7]-s uuriti, kas rikka morfoloogiaga eesti keeles on kestuse prognoosimisel lisaks morfoloogilisele infole abi ka sõnaliigi tundmisest ja süntaktilisest teabest.

[P8]-s keskenduti vajalike tunnuste valikuprintsiipidele tekst-kõne sünteesi kõne ajalise struktuuri modelleerimiseks. Lisaks traditsioonilistele parameetritele, mis kirjeldavad häälikuümbrust ja tema hierarhilist paiknemist lausungis, on segmentaalkestuste prognoosimisel eesti keeles olulised ka sõnade morfoloogilised, süntaktilised ja leksikaalsed tunnused nagu sõnavorm, lauseliige ja sõnaliik. Pauside asukoha prognoosimisel kõnevoos olid tähtsateks tunnusteks sõna kaugus lause algusest ja eelmisest pausist, viimase kõnetakti pikkus ja välde ning kirjavahemärgid või sidesõna tekstis.

1.4. Töös kasutatud mõisted ja kontseptsioonid.

Keele kui märgisüsteemi funktsioneerimise põhieesmärk on tagada mõtete väljendamine ning teabe edastamine ja vastuvõtmine suulise kõne või kirjaliku teksti vahendusel. **Kõne** on keele kui märgisüsteemi kasutamine rääkimisel (suuline kõne), kirjutamisel (kirjalik kõne), mõtlemisel (sisekõne) või muusugusel teatamisel. Kõneoskus ei ole kaasasündinud, vaid omandatakse inimese tegevusega. Inimese bioloogiliste eeldustega antud sünnipärane keelevõime on

loonud aluse keelesüsteemi omandamiseks kõnest ja omandatu kasutamiseks kõnes (Õim 1976).

Keeleline suhtlemine on niisiis mõtete edastamine ja vastuvõtmine kõnesignaalide vahendusel. Arvutid paraku veel mõtelda iseseisvalt ei oska. **Kõnesüntees** või täpsemalt **tekst-kõne süntees** on seadme või arvuti oskus teisendada ortograafilist teksti ortoeepiliseks kõneks ilma inimese osaluseta.

Foneetika uurib keelemärgi väljenduskülge vormistatuna suuliseks kõneks. Foneetika põhiüksus **häälik** on väikseim kuuldeliselt eristatav artikulatoorsete ja akustiliste omadustega määratletav kõnesegment. Samas on häälikul akustilises ruumis väga suur hulk eri variante sõltuvalt häälikulisest ümbrusest sõnas ja konkreetsest kõnelejast. Häälikuerinevuste süstemaatilise taandamise teel saame teada keele **fonoloogilise süsteemi**, mille üksusteks on **foneem**id (Hint 1998). Seega kõnesünteesi sisendis me eeldame teksti või foneemide jada, mis väljundis realiseerub häälikute jadana e **sünteeskõne**na. Kõnetuvastusel on protsess vastupidine, me püüame analüüsil kõnelainest tuvastada häälikute süvastruktuuri e foneemide jada. Kalevi Wiik on tabavalt foneemi ja hääliku vahekorda võrrelnud laskuri olukorraga lasketiirus (Wiik 1991): nii nagu laskuri eesmärgiks on tabada märklaua keset, nõnda püüab kõneleja näiteks erinevates sõnades *sada, tanu, pali* saavutada sama foneemi /a/ sihtväärtust, kuid koartikulatoorsest ümbrusest tingituna on tulemus nagu märklaualgi mitte täpselt sama kvaliteediga häälik vaid lähedaste häälikute kobar. Keele väikseimaid üksusi – **segmentaalfoneem**e – kirjeldatakse nii häälikute kvalitatiivsete omaduste kui ka ajalise mõõtmega seotud parameetriga – **omakestus**ega.

Kõne, so suulise teksti (aga samuti muusika) esituses on oluline teatav korrastus, mis ilmneb häälikutest (foneemidest) pikemas kõnelõigus. See korrastus antakse edasi helisignaali füüsikaliste parameetrite kestuse, põhisageduse ja intensiivsuse muutuste kaudu. See on ala, millega tegeleb **prosoodia**. Füüsikalistest suurustest tuletatud psühhoakustiliste tajuparameetrite pikkuse, kõrguse ja valjuse või nende mitmesuguste kombinatsioonide alusel moodustuvate prosoodiliste tunnustega on kirjeldatavad **suprasegmentaalfoneem**id ehk **prosodeem**id, mis kaasnevad ühe või harilikult mitme segmentaalfoneemiga. Prosodeemide tähendusi eristav võime põhineb mitte niivõrd üksust moodustavate segmentaalfoneemide kvaliteedi erinevustel, kuivõrd kogu üksust iseloomustavate prosoodiliste tunnuste distinktiivsel erinevusel. Sõltuvalt suprasegmentaalse ehk prosoodilise nähtuse olemusest võib prosodeemiga haaratavaks kõnesegmendiks olla kas silp, takt, sõna, sõnaühend või lause. Prosoodiliste nähtuste hulka kuuluvad näiteks sõnarõhk, fraasirõhk, esiletõsterõhk (fokuseeritus), silbitoonid (nt hiina keeles), tonaalsed sõnaaktsendid (nt rootsi keeles), eesti välted, lause intonatsioon jm.

Füüsikaline parameeter **kestus** tähistab igasuguse kõneüksuse (hääliku, silbi, kõnetakti, sõna, fraasi, lause, pausi jms) või selle osa hääldamiseks kuluvat aega. Kestus võib sõltuda vaadeldava üksuse enda (nt hääliku kvaliteedist sõltuv omakestus), aga ka tema naabrite kvalitatiivsetest omadustest ja hulgast, asen-

dist sõnas ja lauses, paljudest muudest morfoloogilistest, süntaktilistest ja paralingvistilistest teguritest (Eek, Meister 2003). Kõneüksuse kestust tajutakse harilikult selle **pikkus**ena (nt lühikese või pika häälikuna, *resp* foneemina).

Kõneüksuse **põhisageduse** (põhitooni, F_o) ja selle muutuse (so erinevaid põhitoonikontuure) tekitab häälekurdude võnkumine heliliste häälikute artikuleerimisel, mida kuulaja tajub helikõrgusena või selle muutusena. Põhitooni kulg fraasis või lauses on aluseks selle fraasi või lause **intonatsioon**ile. Põhitooni kõrgus ja/või selle muutus silbis iseloomustab silbitoone taktis – tonaalseid sõnaaktsente. **Intensiivsus** on kõnelaine energeetiline parameeter, mis väljendab kopsude ja häälekurdude koostoimel tekkivaid õhurõhu erinevusi, aga samuti artikuleerimise pingsusastet, mida kuulaja tõlgendab signaali valjusena.

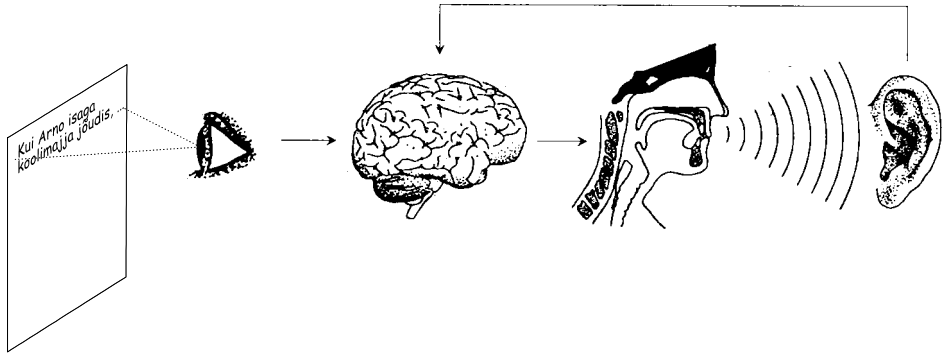
Rõhk on kompleksne hierarhiline prosoodiline nähtus, mida keele fonoloogilisest süsteemist sõltuvalt iseloomustavad erinevad füüsikalised parameetrid (kestus, põhisagedus, intensiivsus, aga ka vokaalide kvaliteet). Madalaima esinemistasandi rõhku nimetatakse **sõnarõhuks**, mis keeliti on kas fonoloogiline (nt inglise ja vene keeles) või afonoloogiline nähtus (nt eesti keele omasõnades on sõnarõhul harilikult piiri markeeriv funktsioon). Pikemates sõnades on mitu rõhku, millest tugevaimat nimetatakse sõna **pearõhuks** ja nõrgemaid **kaasrõhkudeks**. Eesti keele omasõnades langeb pearõhk harilikult sõna esimesele silbile (*resp* taktile). Kõnetakt kaheosalisena koosneb tugevast (s.o rõhulisest) ja nõrgast (s.o rõhuta) silbist. Irdsilbina saab takti kuuluda ka nõrk (rõhuta) kolmas silp, kui see lõpeb lühikese vokaaliga või sõnalõpus ka lühikese konsonandiga. Eesti keele ühesilbilistes sõnades moodustab takti nõrga osa n-ö virtuaalne silp, mis väljendub ühesilbilise sõna lõpu kestuse pikenemises. Kõrgemal tasandil, s.o fraasis või lauses esiletõstetud sõnades langevad mitmesugused esiletõsterõhud enamasti vastava sõna pearõhulisele taktile (Eek, Meister 2004). Eesti keeles väljendab sõnarõhku rõhulise silbi rõhuta silbist kõrgem F_o tippsagedus kõnetaktis (Eek 1987). **Esiletõsterõhk**e eristab tavalisest sõnarõhust pearõhulise takti rõhulise silbi tajutavalt kõrgem F_o sagedus (Asu 2004). Rõhkude vaheldus tekitab kõnerütmi.

Eesti välted kuuluvad prosoodiliste nähtuste hulka. **Välted** on takti piires rõhulise ja rõhuta silbi tuumast koosnevast osast moodustunud ja iseseisvunud distinktiivsed prosoodilised üksused, mille eristatavus sõltub selle taktiosa naaberfoneemide kestuslikest suhetest ning põhitoonikontuuride (ja võib-olla täiendavalt ka intensiivsuskulu, vokaali-konsonandi liitumisviisi ning vokaalide kvalitatiivseteski) erinevustest (Eek, Meister 2004).

2. ÜLEVAADE SÜNTEESI STRATEEGIATEST JA KÕNE AJALISE STRUKTUURI MUDELITEST TEKST-KÕNE SÜNTEESIS

Tekst-kõne süntesaatorite eeskujuks on olnud inimlugemine. Joonisel 1 on esitatud teksti häälega ettelugemise lihtsustatud skeem ja kujutatud inimese füsioloogilised organid, mis on kaasatud lugemisprotsessi.

Inimene omandab lugemisvõime esimesel elukümnendil, edasises elus lugemisoskus areneb ja täieneb. Olles selle võime omandanud, muutub see automaatseks tegevuseks. Vaadeldes lugemist füsioloogia tasandil, näeme, et tegu on väga keeruka protsessiga. Tähemärkide kujutis haaratakse silmade sensorneuronite poolt ja kantakse elektriliste stiimulite vormis inimese ajju, kus see informatsioon töödeldakse ja formeeritakse motoorsete neuronite käsklusteks, mis kannavad hoolt kopsude, häälekurdude ja artikulatsioonilihaste aktiveerimise eest (Holmes 1988). See viib kõne tekitamisele, kusjuures artikulatsiooniprotsessi jälgitakse ja juhitakse pidevalt põhiliselt kuulmisorganitest saabuva informatsiooni põhjal.



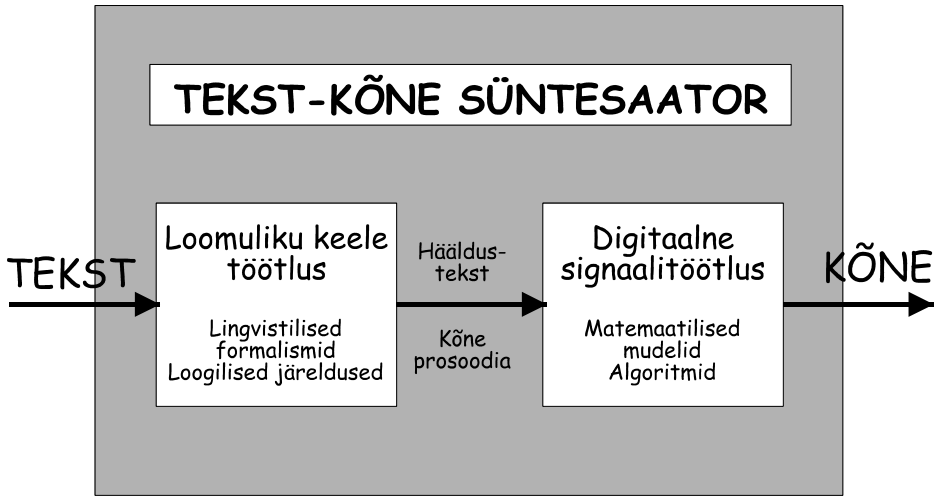
Joonis 1. Andmevoo skemaatiline diagramm illustreerimaks lugemisprotsessi Holmes'i järgi (Holmes 1988).

2.1. Sünteesi strateegiad

Arvutil imiteeritav tekst-kõne süsteem on lihtsustatud mudel füsioloogilisest lugemisprotsessist (joonis 2).

Nii nagu inimlugemine, sisaldab tekst-kõne süntesaator loomuliku keele töötlusmoodulit, mis teisendab sisendteksti hääldustekstiks koos soovitud intonatsiooni ja kõnerütmiga. Digitaalne signaalitöötlusmoodul teisendab sisendis oleva sümbolinformatsiooni loomuliku kõlaga kõneks.

Loomuliku keele töötlusmoodul annab teksti foneetilise kirjelduse ja paneb paika kõne prosoodia. Üldjuhul sisaldab tekstitöötlus keele erinevaid kirjeldustasandeid: foneetikat, fonoloogiat, morfoloogiat, süntaksit ja semantikat.



Joonis 2. Üldistatud tekst-kõne sünteesi mudel.

1960-ndatel aastatel jagunesid kõnesünteesi tehnikad kaheks paradigmaks. Lingaard nimetas neid süsteemi ja signaali meetodiks (Lingaard 1985). Süsteemi meetodit nimetatakse ka artikulatoorseks sünteesiks. Artikulatoorne süntees baseerub kõneloome füsioloogilisel mudelil ja kõnetraktis hääle tekitamise füüsikalisel kirjeldusel. Mõlemad meetodid arenesid sõltumatult, aga kiiremaid reaalseid tulemusi saavutati signaali modelleerimisel, tingituna selle lähenemisviisi sisemisest lihtsusest. Vastupidiselt artikulatoorsele lähenemisele, ei püüagi ta seletada koartikulatsiooni-mõjusid kõneorganite kinemaatika põhjal, vaid lihtsalt kirjeldab vastavaid akustilisi lainekujusid.

Arusaadava ja loomuliku väljundkõne saamiseks on raskuspunkt häälikult häälikule üleminekute ja koartikulatsiooni modelleerimisel. Kõneteaduses on ammu teada, et foneetilised siirded on kõne arusaadavuse seisukohalt mitte vähem olulised, kui statsionaarsed osad (Liberman 1959). Foneetiliste siirete arvestamine sünteesis on saavutatav kahel viisil: otseselt – reeglitejada vormis, mis formaalselt kirjeldab foneemide mõju üksteisele; kaudselt – salvestades foneetilised siirded ja seega koartikulatsioonilised mõjustused kõnesegmentide andmebaasi ja kasutades neid sünteesil lõplike akustiliste üksustena foneemide asemel.

Mainitud kahest alternatiivist on arenenud kaks põhilist tekst-kõne süsteemi liiki – reegelsüntees ja ahelsüntees. Mõlemal on oma sünteesifilosoofia.

Reeglitel põhinevad süntesaatorid on soositud foneetikute ja fonoloogide seas, sest neid saab kasutada häälduse mehhanismide uurimiseks. Kõige laialdasemat kasutust on leidnud nn Klatt'i süntesaator (Klatt 1980), sest artikulatoorsete parameetrite ja Klati mudeli sisendite vahelise seose tõttu on võimalik seda süntesaatorit kasutada kõnefüsioloogia uurimisel. Erinevalt reegelsünteesist on kõneüksuste ühendamisel põhinevatel süntesaatoritel väga vähe informatsiooni käsitletavate andmete kohta. Enamik infost sisaldub segmentides, mida jadas ühendatakse.

Ahelsüntees eeldab, et artikuleeritud kõnevoog ei ole lihtne ritta seatud häälikute jada. Pigem koosneb kõne pidevalt kattuvatest üleminekutest ühelt häälikult teisele. Regressiivse koartikulatsiooni tõttu eelnev segment sisaldab järgmise kõnehääliku tunnuseid. Difoonid [1] on ahelsünteesil enimkasutatud kõneühikud, kuna suvalise teksti alusel kõne genereerimiseks on vaja suhteliselt väikest arvu difoone. Eesti keele difoonide andmebaas sisaldab ligikaudu 1900 difooni. Kui tavalisel tekst-kõne difoonsünteesil on kõne andmebaasis täpselt üks häälikult-häälikule üleminek, siis korpuspõhisel sünteesil on kogu korpus sünteesi akustiliseks baasiks. Difoonid on elementaarühikuna kasutusel ka muutuva pikkusega kõneüksuste korpuspõhisel sünteesil (Clark jt 2007). Kõneüksuste valikualgoritmid alustavad otsinguid fonoloogilise puu kõrgematelt tasanditelt (fraas, sõna, kõnetakt) eelistades sünteesil võimalikult pikemaid kõnelõike.

Käesolevas töös on kõne ajalise struktuuri modelleerimisel eelkõige orienteeritud nii üksikutel difoonidel põhinevale tekst-kõne ahelsünteesile (Mihkla, Meister 2002) kui ka korpuspõhisele ühikute valiku sünteesisüsteemile (Mihkla jt 2007). Kuna difoonid kätkevad endas naaberhäälikute üleminekut, siis on mõttekas kõne ajalise struktuuri elementidena käsitleda häälikute ja pauside segmentaalseid kestusi.

2.2. Kõne ajaline regulatsioon

Kõne ajalises juhtimises on olnud kolm põhilist lähenemisviisi – moora-ajastus rütm, mida on rakendatud nt jaapani keeles, silbi-ajastus rütm, mis on omane eelkõige prantsuse ja hispaania keelele ning rõhu-ajastus rütm, mida on tuvastatud ja rakendatud paljude indo-euroopa keelte ajalises regulatsioonis.

Jaapani keeles on mooraisokrooniat täheldatud ajalise kitsendusena just vokaalide kestuse juhtimisel. Negatiivne korrelatsioon on tuvastatud vokaalide kestuse ja naaberkonsonandi kestuse vahel. Vokaali kestuse kompensatsioon on rohkem mõjutatud vokaalile eelneva konsonandi kestusest ja seda vaadeldakse moora-ajastuse akustilise ilminguna. Statistilise analüüsi kaudu on leidnud

[1] Difoonid algavad mingi hääliku stabiilse osa keskelt ja lõpevad järgmise hääliku stabiilses osas.

kinnitust, et selline kompensatsioon leiab aset moora üksustes aga mitte silbis (Sagisaka 2003). Moorameetrikat on edukalt rakendatud ka eesti keele fonoloogias. Arvo Eek tõlgendas eesti sõnaprosoodias takti piires välteid kui mooraisokroonia ilmingut, kus kestuste taktisisene jaotus määrab välte (Eek, Meister 2004:336–357).

Silbi-ajastus keeles eeldatakse, et iga silp, mida hääldatakse on ligikaudu võrdse kestusega, ehkki silbi tegelik kestus sõltub situatsioonist ja kontekstist. Hispaania ja prantsuse keelt on klassifitseeritud silbiajastuskeelteks, kuigi päris kindlat nõustumist selles osas ei ole (nt Wenk, Wioland 1982). Kui kõneleja kordab ühte ja sama lauset mitu korda samas kõnetempos, siis naaberhäälikute kestused näitavad tugevat negatiivset korrelatsiooni, st iga üksikhääliku kestuse variatsioon kompenseeritakse naaberhäälikute kestusega. Seega artikulatsiooni kestuslik regulatsioon peab haarama foneemist kõrgema, näiteks silbi tasandi (Huggins 1968). Silbiajastuse hüpoteesi rakendasid Campbell ja Isard kõrgemate ja madalamate tasandite seoste statistiliseks modelleerimiseks (Campbell, Isard 1991).

Rõhuajastus rütmiga keeltes võivad silbid kestuselt olla erinevad, aga kahe järjestikuse rõhulise silbi vahelise lõigu kestus on keskmiselt konstantne. Isokrooniat on paljudes keeltes kaua ja põhjalikult uuritud, aga ühtset seisukohta kõne ajalise regulatsiooni ja tema akustiliste tunnuste kohta pole veel esitatud. Ilse Lehiste tuli ulatuslikus ülevaates (Lehiste 1977) isokroonia ja kõne rütmilisuse tõendite kohta järeldusele, et inglise keeles puuduvad kõne rütmilisusega seotud otsesed akustilised korrelaadid. Ilmselt peab nõustuma Thierry Dutoit väitega, et nö „puhtaid" keeli, mis täpselt vastaksid eespool toodud ühele või teisele rütmimudelile, ei olegi olemas ja pigem on adekvaatne öelda, et keeltes on vaid tendents isokrooniale (Dutoit 1997). Eesti vältesüsteemi käsitlevates hiljutistes töödes peetakse sobivaks kirjeldada välteid taktiisokroonia kontekstis (Wiik 1991; Eek, Meister 2003).

2007.a. foneetikateaduste kongressil Saarbrückenis oli kõne ajastusele pühendatud eri istungjärk, kus eri keelte (inglise, jaapani, brasiilia portugali ja prantsuse) uurijad käsitlesid kõne rütmilisuse mehhanisme. Ehkki päris ühist lähenemist ei olnud, oli paljude uurijate tähelepanu fokuseeritud vokaali alguste (vowel onset) eri aspektidele kõne ajalises struktuuris (Keller, Port 2007). Helilisuse algused on tänu nende silmapaistvusele tajumisel andnud võtme silbi ajalise ülesehituse uurimiseks. Vokaalialgused etendavad otsustavat rolli kõnesünteesi kvaliteedi loomulikustamisel ja nad sisaldavad kõne tajumisel olulisi parameetreid (Keller 2007). Huvitaval kombel on kongressi istungjärgul kirjeldatud uus lähenemisviis väga sarnane eesti väldete kõnetakti teooriaga, kus olulist rolli mängivad just rõhulise silbi riimi ning rõhuta silbituuma kestussuhted[2].

[2] Välde kõnetaktis on defineeritud $\sigma_{rõhuline}$(nucleus+[coda]) / $\sigma_{rõhuta}$(nucleus).

Eesti keel on ilmselt rõhu-ajastus rütmiga. Antud töös lähtutakse kõne kestuslikul modelleerimisel eesti keele silbi- ja taktiehituse põhijooni arvestavast välte ja rõhu käsitlusest.

2.3. Statistilised meetodid prosoodia modelleerimisel

Teadus järgneb tehnoloogiale ja piirangud tehnoloogias mõnikord kitsendavad teaduslikku vaadet (Campbell 2000). Veel kakskümmend aastat tagasi, kui kestusi mõõdeti ostsillogrammidelt ja spektrogrammidelt, oli uuritava kõnelõigu kestuse piiranguks paberi mõõtmed, millele sai trükkida. Sellest tulenes, et enamik andmeid põhines varasemates töödes sõnadel või fraasidel, mis olid esitatud lühikestes raamlausetes. Et analüüsi maht olid piiratud, siis keskenduti eelkõige nn „laboratoorsele kõnele", milles segmentide kestused võivad erineda sidusast kõnest mõõdetutega märgatavalt (Campbell 2000). Hiljem, kui tekkis võimalus automaatselt analüüsida ja töödelda kõne suuremahulisi andmebaase, hakati kõne ajalist struktuuri uurima sidusa kõne baasil. Teine põhjus kõne statistilisele modelleerimisele üleminekuks kätkes reeglipõhistes prosoodiasüsteemides endis.

Reeglipõhised kõne ajalise struktuuri juhtimismudelid määrasid segmentide kestuste väärtusi enamiku juhtude jaoks, paraku ilmnesid mõnikord ka tõsised vead. Need vead olid sageli põhjustatud sellest, et samaaegselt püüti rakendada sõltumatult tuletatud reegleid. Kui aga suured kõne andmebaasid muutusid kättesaadavaks, hakati neid kasutama, et ära hoida reeglipõhise modelleerimise vigu ning täpsemalt määrata kestusi, rakendades statistilisi protseduure segmentaalsete kestuste ennustamiseks.

Väljakutse kestusi prognoosida on atraktiivne nii matemaatikutele kui lingvistidele. Esimeseks pioneeriks kestuste statistilise modelleerimise vallas peetakse Michael Riley't, kes 1989. a. kirjeldas CART-meetodi (classification and regression trees) rakendamist segmentaalsete kestuste prognoosimiseks (Riley 1989). CART genereerib andmete põhjal kahendpuu, jagades neid rekursiivselt osadeks ja minimeerides vea variatiivsust. Sellest ajast peale on ilmunud suur hulk töid mitmesuguste statistiliste meetodite kasutamisest kõneüksuste kestuste ennustamiseks paljude keelte kohta. Nick Campbell võttis esimesena kasutusele närvivõrgud silbi kestuste arvutamiseks konteksti põhjal. Jaapanlased on põhiliselt jäänud truuks regressioonimudelite kasutamisele prognoosil (Kaiki jt 1992; Sagisaka 2003). Vaatamata sellele, millist konkreetset prognoositehnikat rakendatakse, on statistilisel modelleerimisel mitu eelist reeglipõhiste süsteemide ees.

Esimeseks eeliseks on täpsus ja selgus modelleerimisel. Statistiline optimeerimine välistab suured vead, mis on näiteks põhjustatud kestuste juhtimisreeglite ettenägematult halvast kombinatsioonist. Veelgi enam, statistilised

tehnikad teevad võimalikuks analüüsida väikesi, varjatud, kuid olulisi erinevusi [P7]. Suurte vigade kahandamine parandab kindlasti sünteeskõne loomulikkust ja täppisanalüüsi võimalused annavad hea pildi regulatsioonimudelitest foneetikas (Sagisaka 2003).

Teine eelis on teaduslikus baasis, mis on korpuspõhise modelleerimise aluseks. Reeglipõhises sünteesis ei ole selget andmete kirjeldust, juhtimisalgoritme ja veamõõtmise võimalust paljudel juhtudel. Korpuspõhisel statistilisel modelleerimisel saame teada kestuste regulatsiooni täpsuse piire ja infot selle parandamiseks, muutes kas korpust, juhtimisalgoritme või vea mõõtmisi. Seega oleme me saanud teadusliku süstemaatilise meetodi, et pakkuda välja vea analüüsi tulemusi tagasisidena empiirilise reegelpõhise rakenduse arendamiseks. Loodetakse, et selline korpuspõhine statistiline lähenemine saab enamlevinuks foneetika teadustes, kus iga teooriat on tavaliselt testitud erinevates tingimustes ja erinevatel andmetel ja mõõtmistel (Sagisaka 2003).

Käesolevas töös rakendatakse erinevaid statistilisi meetodeid (lineaarne ja logistiline regressioon, närvivõrgud ja CART) kõne ajalise struktuuri modelleerimiseks teksti- ja kõnekorpuste baasil.

3. EESTIKEELSE KÕNE AJALISE STRUKTUURI UURIMUSED JA MODELLEERIMINE

Eestikeelse kõne ajalise struktuuri kohta on ilmunud hulk töid, milles on püütud kõne kestuslikku struktuuri lihtsalt kirjeldada või kõneprosoodia nähtusi eksperimentaal-foneetika mõõtmiste tulemustele toetudes terviklikult käsitleda.

Eesti keele prosoodia arenguloost on teinud Taeve Särg põhjaliku ülevaate oma doktoritöös (Särg 2005): „17.–19. sajandil kirjutatud keele- ja luulealaste tööde põhjal alles teadvustati eesti keeles sõnade tähendust eristavaid ning keele ja rahvalaulu vormi seisukohalt olulisi prosoodilisi tunnuseid." Tol ajal mõjutasid prosoodia kirjeldamist indoeuroopa keelte põhjal väljakujunenud arusaamad, mille suur vastuolu eesti keelega seisnes selles, et kuni 19. sajandi lõpuni ei tehtud neis teoreetilist vahet rõhul ja kestusel (Preminger, Brogan 1993).

Kui 20. sajandi esimese poole foneetikaalased kirjutised ja foneetika ülevaated toetusid eesti keele õigehäälduse kirjeldamisele, siis sajandi teise poole foneetikauurimused tuginevad juba eksperimentaalfoneetika aparatuuri rakendamisele ja hiljem arvutite laialdasele kasutamisele. Kaasaegsest objektiivsetele mõõtmistele tuginevast eestikeelse kõne ajalise struktuuri uurimisest saame rääkida alates 1960-ndatest aastatest (Lehiste 1960; Liiv 1961; jt). Järgnevas vaatleme neid eestikeelse kõne kestuslikku struktuuri käsitlevaid töid, mis põhinevad eksperimentaalfoneetikal.

Kõne ajalise struktuuri käsitlemisel on enam tähelepanu pööratud kvantiteedisüsteemile (so väldetele) kui eri häälikute segmentaalkestustele. Eesti prosoodias tunnustatakse kestuse kontrastiivset kasutust. Kontrastiivsed välted eesti keeles on lühike, pikk ja ülipikk, vastavalt tähistatuna Q1, Q2, Q3. Väldete abil saab eesti keeles leksikaalseid ja grammatilisi erinevusi väljendada ainuüksi kvantiteedi abil, muutmata sõna häälikulist koosseisu (nt *jama, jaama* Gen, *jaama* Part; *suga, suka* Gen, *sukka* Part).

Eesti keeles on 9 vokaalfoneemi ja 17 konsonantfoneemi. Kõik vokaalid võivad esineda kolmes kontrastiivses kvantiteedis sõna esimeses silbis, samamoodi võivad peaaegu kõik konsonandid esineda kolmes kontrastiivses vältes esimese ja teise silbi piiril. Ilse Lehiste mõõtmiste alusel on lahtiste esimeste silpide vokaalide kestused kolmes vältes keskmiselt 110, 180 ja 230 ms, ligikaudse suhtega 2:3:4 (Lehiste 1960). Lingvistilise kvantiteedi ehk välte tajumisel pole niivõrd tähtis häälikukestus, vaid ülalmärgitud taktisegmentide kestussuhted.

Tabel 1. Uurijate poolt mõõdetud rõhulise ja rõhuta silbi kestussuhteid.

	Q1	Q2	Q3
Lehiste 1960	0.7	1.5	2.0
Liiv 1961	0.7	1.6	2.6
Eek 1974	0.7	2.0	3.9
Krull 1991,1992	0.5–0.7	1.2–2.1	2.2–2.9
Alumäe 2007[3]	0.6–1.0	1.5–2.6	2.1–4.0

Eesti keele prosoodilist süsteemi vaadeldakse hierarhilisena: segment (foneem), silp, kõnetakt, sõna, fraas, lause. Siin on peamine küsimus, millisel hierarhiatasandil on kvantiteedinähtusi kõige otstarbekam kirjeldada. Kui kunagi väljapakutud häälikuvälte teooria pole leidnud poolehoidu, siis enamik uurijaid on välteid määratlenud kas silbisuuruste üksustena (Hint 1997; Viitso 2003) või määranud vältehaardeks rõhulisest ja rõhuta silbist koosneva takti (Wiik 1985; Eek, Meister 1997; Lehiste 1997; Ross, Lehiste 2001). Kestuse mõõtmised on näidanud, et välteid iseloomustab kõnetaktis rõhulise ja rõhuta silbi teatav kestussuhe (Lehiste 1960, Eek, Meister 1997). Tabelis 1 on toodud eri uurijate poolt kõnetaktis mõõdetud rõhuliste ja rõhuta silpide kestussuhteid.

Kui varasemad kvantiteedi uurimused põhinesid suuresti nn „laboratoorsel kõnel" (isoleeritud sõnad, sõnad konstrueeritud raamlausetes või isoleeritud laused), siis Diana Krull tõestas, et need iseloomulikud suhted säilivad ka spontaanses kõnes (Krull 1997).

Arvo Eek ja Einar Meister pakuvad silpide kestussuhete asemele välja uusi foneetilisi korrelaate väldete liigitamisel tempokorpuses tehtud uuringute põhjal. Silbi- ja taktivälte teooria vastandamise asemel nad tõdevad: „On tarbetu rääkida eraldi silbi- ja taktivältest, eriti kui nn silbivältegi kolmikvastandus ilmneb takti piires ja kui väldet ei tunta ära rõhulise silbi vaid takti foneetiliste omaduste kaasabil. Seetõttu on mõistlikum kõnelda lihtsalt välteist." (Eek Meister 2003)

Ehkki kestussuhted mängivad väldete tajumisel olulist rolli, on näiteks Q2 ja Q3 eristamisel tähtis osa ka põhitoonil (Lehiste 1960; Liiv, Remmel 1975; Eek 1987). Sageli tuleb kõneprosoodias kõne kestuslikku struktuuri käsitleda koos põhitooni ja intensiivsusega. Eesti keele lause intonatsiooni väga ulatuslik ja põhjalik käsitlus on esitatud Eva Liina Asu doktoritöös (Asu 2004).

Häälikute kestuste prognoosimisel on oluline teada häälikute omakestusi ja mõjutusi naaberfoneemidest. Omakestusi ja häälikute omavahelisi mõjutusi on uuritud paljudes keeltes. Keele need universaalsed nähtused ilmnevad ka eesti keeles. Vokaalide omakestuste esimesed mõõtmised toimusid ligi pool sajandit tagasi (Liiv 1961). Mitmetes hilisemates eesti keele häälikute sisemiste mikroprosoodiliste variatsioonide uurimustes on tõdetud, et lühikeste madalate

[3] Suhted on arvutatud automaatselt segmenteeritud sidusa kõne põhjal (vt http://keeletehnoloogia.cs.ut.ee/konverents/slaidid/alumae.pdf)

vokaalide kestused on 10–15 ms pikemad kõrgete vokaalide kestustest (Eek, Meister 2003:836; Meister, Werner 2006:111). Naaberfoneemide omavaheliste mõjutuste alla kuuluvad niisugused nähtused nagu konsonantide lühenemine konsonantühendites ning eriti siis kui nende naabriteks on helitud konsonandid (Eek, Meister 2004:267).

Pause ja lõpupikendusi on eestikeelses kõnes uuritud põgusalt või riivamisi teiste ülesannete kontekstis. Ilse Lehiste kontrollis kas lõpupikendused on korrelatsioonis järgnevate pauside pikkustega ja tuvastas väga nõrga seose (Lehiste 1981). Diana Krull uuris pausieelseid pikendusi dialoogkõnes kahesilbilistes sõnades väldete kontekstis (Krull 1997). Arvo Eek ja Einar Meister mõõtsid lauselõpu pikendusi tempokorpuse baasil (Eek, Meister 2003). Aga ka neil oli vaatluse all vaid kindla struktuuriga sõnad ja põhitähelepanu keskendus väldete tunnustele. Seetõttu tekkis vajadus eestikeelse tekst-kõne sünteesi jaoks mõõta pause ja lõpupikendusi sidusas kõnes.

Üheks esimeseks uurijaks, kes püüdis reeglitejada vormis eesti keele välteid modelleerida, oli Kalevi Wiik. Ta esitas Arvo Eegi väldete mõõtmisandmeid moorameetrika süsteemis ja tuletas sellel alusel sünteesireeglid (Wiik 1985). Eelmise sajandi kaheksakümnendatel aastatel töötati Küberneetika Instituudis välja mitmeid parameetriliste kõnesüntesaatorite prototüüpe. Nende süntesaatorite tarvis loodi ka reeglipõhised prosoodiamudelid, mis juhtisid sünteeskõne ajalist struktuuri ja intonatsiooni (Meister 1991; Siil 1991).

Aastatel 1997–2002 loodi eesti keele tekst-kõne sünteesi prototüüp. Süntesaator põhines difoonidel ja reeglipõhisel prosoodiamudelil (Mihkla jt 2000). Kõnelaine ajalise struktuuri reeglistamisel arvestati vokaalide omakestusi, väldete kestussuhteid kõnetaktis ja eesti keele rõhu käsitluse ja silbiehituse põhijooni. Kõne ajalise struktuuri mudel sisaldab mitmeid kestuste tabeleid ja suure hulga reegleid, mis juhivad häälikute kestusi sõltuvalt kontekstist. Pauside ja piiripikenduste kestuste väärtusi ei modelleerita, nad lisatakse kõnevoosse konstantsete suurustena.

Eesti keele kõne ajalist struktuuri statistiliste tehnikatega teadaolevalt varem modelleeritud ei ole.

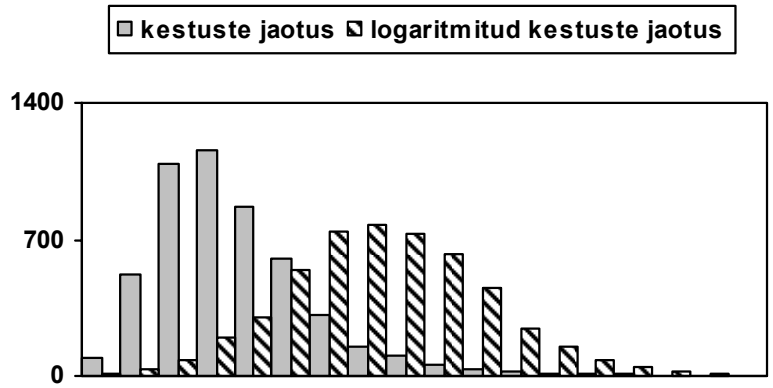
4. ANDMED

Uurimuse eesmärgiks on analüüsida ja modelleerida häälikute ja pauside kestusi sidusas kõnes eestikeelse tekst-kõne sünteesi tarbeks. Seetõttu lähtematerjaliks valiti diktorite poolt etteloetud erinevat tüüpi tekstid. Teksti ja kõne üks-ühese vastavuse põhjal saab prosoodia sümbolesituselt üle minna akustilisele ning samuti tuvastada, kas ja kuivõrd on teksti süntaktiline liigendus seotud kõne prosoodilise liigendusega.

Lähtematerjaliks võeti kõnelõigud näitleja poolt ette loetud kriminaalloo CD-versioonist (Stout 2003), kõnelõigud ning tekstid Eesti Raadio pikematest diktorite loetud uudistest ja kõnelõigud eesti foneetilisest andmebaasist BABEL (Eek, Meister 1999).

Kokku oli analüüsi all 66 kõnelõiku, 27 diktori (14 mehe ja 13 naise) esituses. Diktorid lugesid erinevaid kõnelõike, vaid Babeli andmebaasi salvestuste korral lugesid ühte ja sama teksti 2–3 diktorit. Kogu kõnematerjal segmenteeriti käsitsi häälikuteks ja pausideks. Et eesti keele foneetilise andmebaasi kõnelõigud olid juba kõneüksusteks jaotatud, siis ülejäänud materjali märgendusel kasutati sedasama foneetilist transkriptsioonisüsteemi (Eek, Meister 1999). Kogu kõnematerjali mahuks oli 46 minutit kõnet, millest kõige mahukam materjal 9.25 minutit kõnet oli naisraadiodiktori esituses.

On hästi teada fakt, et segmentaalkestused järgivad normaaljaotust logaritmilises skaalas, mistõttu enamikes modelleerimiseksperimentides [P1], [P2], [P3], [P5], [P6], [P7] ja [P8] kasutati funktsioonitunnusena logaritmitud kestust (joonis 3). Sisendid e argumenttunnused genereeriti etteloetud tekstide põhjal. Liitsõnapiiri, 3. välte ja palatalisatsiooni määramiseks kasutati tekst-kõne sünteesi jaoks loodud lingvistilise töötluse moodulit (Kaalep, Vaino 2001). Lause süntaktilist analüüsi [P4] ja [P7] uurimuste jaoks tegid käsitsi vastavalt Krista Kerge ja Katre Õim. [P7] sõnade morfoloogilise ja sõnaliigi info tuvastamiseks kasutati Eesti Keele Instituudis väljatöötatud meetodeid (Viks 2000).



Joonis 3. Häälikute kestuste ja logaritmiliste kestuste jaotused meesdiktori andmete põhjal.

5. MEETODID

5.1. Töös kasutatud meetodid ja statistilise modelleerimise mõisted

Kõne ajalise struktuuri modelleerimisel kasutati sisendteksti põhjal genereeritud muutujate väärtuste alusel järgmisi statistilisi meetodeid: lineaarset regressiooni ([P1], [P2], [P3], [P4], [P5], [P6], [P7] ja [P8]); logistilist regressiooni ([P3], [P5] ja [P8]); klassifikatsiooni ja regressioonipuid ([P5], [P6] ja [P8]); närvivõrke ([P5], [P6], [P7] ja [P8]).

Kõigi nende statistiliste meetodite kohta on olemas suurepäraseid tutvustusi ja käsiraamatuid, näiteks klassifikatsioonist ja regressioonipuudest (Breiman jt 1984), närvivõrkudest (Gurney 1997), lineaarsest regressioonist (Weisberg 1985) ja logistilisest regressioonist (Hosmer, Lemeshow 2000). Enne, kui minna meetodite rakenduste ja võrdluse juurde, täpsustame väitekirjas kasutatud termineid:

Muutuja – muutuva suuruse sümbol, milles sisalduv informatsioon võib olla kas numbrilises või sümbolvormis;

Sisendid e **argumenttunnused** – muutujad, mille põhjal prognoositakse väljundit (käesolevas töös eeldatakse, et argumenttunnused on determineeritud ja nad moodustavad argumenttunnustevektori $X=(x_1, x_2,\ldots,x_p)$.);

Väljund e **funktsioonitunnus** – muutuja, mille väärtus arvutatakse sisendite põhjal;

Mudel – võrrandite või algoritmide hulk, mille alusel arvutatakse väljundväärtus sisenditest;

Kaalud – numbrilised väärtused, mida kasutatakse mudelis;

Parameetrid – kaalude optimaalsed väärtused mudelis;

Treenimine – kaalude optimaalsete väärtuste määramise protsess mudelis või puustruktuurilise mudeli korral optimaalsete hargnemismuutujate ja –väärtuste valik;

Treenimisandmed – sisend-väljundandmed, mida kasutatakse kaalude määramiseks treenimisel;

Testandmed – sisend-väljundandmed, mida ei kasutata treenimisel;

Valideerimisandmed – sisend-väljundandmed, mida kasutatakse kaudselt treenimise ajal mudeli valikul või treenimise peatamisel;

Kategoriaalne muutuja – muutuja, millel on limiteeritud võimalike väärtuste hulk;

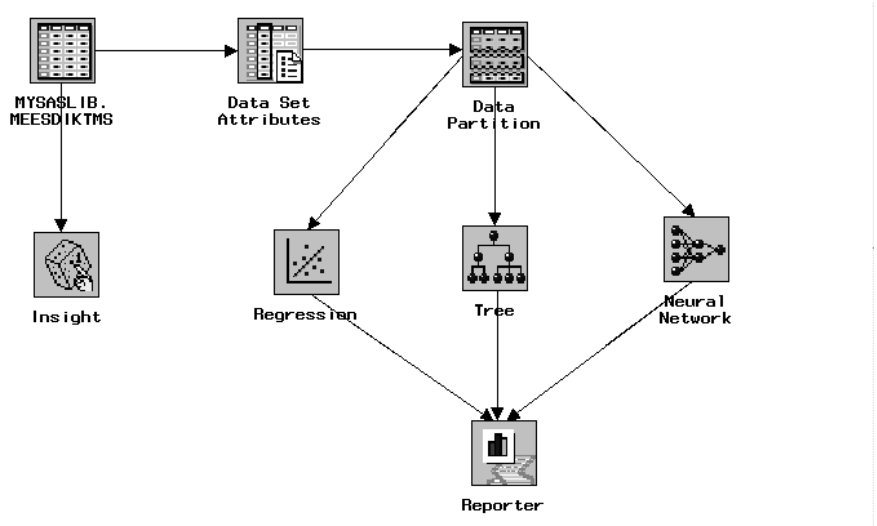
Nominaalne muutuja – numbriline või sümbolkujul kategoriaalne muutuja, milles kategooriad on järjestamata;

Ordinaarne muutuja – numbriline või sümbolkujul kategoriaalne muutuja, milles kategooriad on järjestatud;

Intervallmuutuja – numbriline muutuja, mille puhul väärtuste erinevused on informatiivsed;
Binaarne muutuja – muutuja, millel on vaid kaks erinevat väärtust.

5.2. Töös kasutatud statistikaprogrammid

Esimesed segmentaalkestuste prognoosimised toimusid MS Excel keskkonnas, rakendades lisandmooduli Analysis ToolPaki koosseisus olevat regressioon-analüüsi tööriista ([P1], [P2]). Järgmiseks töövahendiks sai kasutusele võetud statistikaprogrammipakett SYSTAT 11. Selle programmiga sai statistilisel modelleerimisel kasutada nii mitmest lineaarset regressiooni, regressioonipuid kui ka logistilist regressiooni pauside asukoha määramiseks ([P3], [P4], [P5]). Doktorikooli raames osutus võimalikuks kasutada TÜ Rakendusstatistika Instituudi vahendusel statistikaprogrammi SAS 9.1 litsentsi. Töö programmi Enterprise Miner keskkonnas oli statistiliseks modelleerimiseks kõige mugavam, sest samaaegselt sai rakendada erinevaid meetodeid ning võrrelda mudelite sobivust ja eri meetodite tulemusi ([P6], [P7], [P8]). Programmi oli käepärasem kasutada seetõttu, et SAS keskkonnas pole vaja sisendandmeid eelnevalt töödelda (nt teisendada katogoriaalsed muutujad binaarsete pseudomuutujate hulgaks), vaid see toimub automaatselt. Joonisel 4 on kujutatud tüüpiline töös kasutatud andmevooskeem SAS Enterprise Miner keskkonnas.



Joonis 4. SAS Enterprise Miner töökeskkond kõne ajalise struktuuri modelleerimiseks.

Andmevoo moodulite kirjeldus:

MYSASLIB.MEESDIKTMS	– sisendandmed meesraadiodiktori kohta
Insight	– hea tööriist andmetest ülevaate saamiseks muutujate kaupa, selle abil on võimalik tuvastada vigaseid või puuduvaid andmeid
Data Set Attributes	– moodul muutujate funktsiooni määratlemiseks mudelis (st milline on sõltuv muutuja e funktsioonitunnus ja millised mudeli sisendid e argumenttunnused)
Data Partition	– sisendandmete jaotus treening-, valideerimis- ja testandmeteks
Regression	– regressioonanalüüsi moodul
Tree	– otsustuspuude moodul
Neural Network	– närvivõrkude moodul
Reporter	– tulemuste esitlusmoodul

6. TULEMUSED

6.1. Pauside ning piiripikenduste kestuste ja nende asukoha analüüs sidusas kõnes

Et tehiskõne tunduks inimkõrvale loomulik, peaks ta sisaldama loomuliku kõlaga intonatsiooni, rütmi ja rõhuasetust. Ehk täpsemalt, tekst-kõne süsteem peab olema võimeline genereerima selliseid häälikute ja pauside kestusi ning põhitooni väärtusi, mis ei erine oluliselt vastavatest väärtustest reaalses kõnes. (Zellner 1994). Foneetikas ja fonoloogias on pausidele seni suhteliselt vähe tähelepanu osutatud. Suulise kõne lingvistilistes uurimustes on kõneüksustena käsitletud häälikuid, silpe, kõnetakte, sõnu ja fraase põhiliselt isoleeritud lause koosseisus. Lausesiseselt on aga pause raske käsitleda toimivate kõneüksustena, mis võibki olla peapõhjuseks nende lingvistilis-foneetilisele tähtsusetusele (Tseng 2002). Viimasel kümnendil, kui kõnekorpusi hakati laialdaselt kasutama foneetilises uurimistöös, on pausidele kui kõneprosoodia olulisele tunnusele järjest enam tähelepanu pööratud.

Väitekirjas on pause analüüsitud käsikäes segmentaalkestustega ([P1], [P2], [P8]). [P4]-s käsitletakse pause ja kõnehingamist prosoodilise rühma piire markeerivate üksustena. Artiklid [P3] ja [P5] on pühendatud pauside ja piiripikenduste analüüsile ja pauside kestuste ja nende asukoha modelleerimisele kõnevoos. Kui [P3]-s kasutatakse modelleerimisel vaid lineaarset ja logistilist regressiooni, siis pause kokkuvõtvas artiklis [P5] modelleeritakse pauside kestusi veel CART-meetodil ja närvivõrkudega.

Et [P5]-s näite 1 allkiri on artiklis puudulik, siis toome siinkohal selle uuesti (joonis 5) iseloomustamaks pauside paiknemist eestikeelses kõnevoos. Võrdlusena on joonise vasakus veerus etteloetud tekst ja paremal vastava kõnevoo lihtsustatud esitlus – pausid grafeemijadas. Näeme, et teksti struktuur on oluliselt rangem – üldjuhul on iga sõna lõpus tühik ja iga lause lõpus kirjavahemärk. Kõnes võib iga inimene teksti küllalt vabalt interpreteerida: sõnadevahelised pausid paiknevad sõnadegrupi või prosoodilise fraasi järel, aga prosoodilised fraasid ei pruugi kokku langeda süntaktiliste fraasidega ja lõpupikendustel on tendents paikneda prosoodilise fraasi lõpus, aga mitte alati. Osa joonisel 5 allajoonitud pikendatud kõnetakte on seotud fokuseerimisega (nt fraasis *veetlevate noorte naiste seltskonnas* on esile tõstetud sõna *naiste* kõnetakti pikendusega).

Talle meeldis nendega uhkustada – kui need teie omad oleksid, meeldiks see teilegi –, aga mitte sellepärast ei seganud ta vahele. Ta tahtis paari kirja dikteerida ja ta arvas, et kui ma missis Hazeni üles orhideesid vaatama viin, siis ei tea keegi, millal me sealt alla tuleme. Aastaid tagasi jõudis ta ebapiisavatele tõenditele tuginedes otsusele, et ma kaotan veetlevate noorte naiste seltskonnas ajataju, ja kui tema kord midagi otsustab, siis on see otsustatud.

Tallemeeldisnendegauhkustada**P**kuinee dteieomadoleksidmeeldikssee teilegi**P**ag amitteselleparasteiseganudtavahele**P**Tat ahtis**P**paarikirjadikteeridajataarvasetkui mamissisHazeniülesorhideesidvaatama viinsiiseiteakeegi**P**millal**P**mesealtallatul eme**P**Aastaidtagasijõudista**P**ebapiisavat eletõenditeletuginedesotsuseleetmakaot anveetlevatenoortenaisteseltskonnasajat aju**P**jakuitemakordmidagiotsustabsiison seeotsustatud**P**

Joonis 5. Etteloetud teksti struktuur versus pausid kõnevoos. Vasakul veerus etteloetud tekst ja paremal veerus kõnevoos olevad pausid (**P** – sõnadevahelised pausid, allajoonitud grafeemid – pikendatud kõnetaktid).

Artiklites [P3] ja [P5] analüüsiti esmalt neid pause ja lõpupikendusi kõnes, mis olid seotud kirjavahemärkide ja sidesõnadega. Selleks mõõdeti etteloetud tekstide kõnelainetest pauside kestused ja arvutati kõnetakti pikendused. Kõnetakti pikenduste arvutamiseks summeeriti kõnetakti moodustavate häälikute kestused ja võrreldi saadud summeeritud kestust antud taktistruktuuri keskmise kestusega konkreetse diktori kõnes. Lisaks struktuurile arvestati ka taktiväldet. Juhul, kui mingi taktistruktuur osutus antud tekstis unikaalseks (nt CVCCC-CV sõna *'korstna'*) struktuuriks, siis võrreldi tema kestust mingi sarnase kõnetakti struktuuriga (nt CVCC-CV sõna *'kordse',* lahutades ‚korstna' hääliku kestuste summast konsonantühendi ühe komponendi kestuse).

Töödes [P3] ja [P5] on toodud tabelis 1 pauside ja lõpupikenduste keskmised kestused 27 diktori kõnes. Tabelitest on näha, et isegi keskmiste väärtuste variatiivsus on väga suur. Huvitav on siiski märkida, et meeste ja naiste pauside üldkeskmised erinevad kestustelt üksteisest vaid 10% piires. Üldkeskmiste visuaalse vaatluse põhjal võib arvata, et normaalse kõnetempoga etteloetud teksti puhul on pausid kestuse poolest eristatavad. Valimite statistiline analüüs kinnitab seda väidet. Fraasi-, lause- ja lõigulõpu pausid on kõnes kestuselt eristatavad. Analüüsides Studenti t-testiga taktipikenduste andmeid tuli jääda nullhüpoteesi juurde: kõnetakti pikendused olid ühesuguse keskväärtusega valimitest.

Teise sammuna oli vaatluse all, kas ja kuivõrd on kõne prosoodiline liigendus korrelatsioonis teksti süntaktilise liigendusega seal, kus viimast tähtistavad kirjavahemärgid ja sidesõnad. Artiklite [P3] ja [P5] tabeli 2 põhjal, on kõnes paus alati iga lõigu lõpus ja peaaegu iga lause lõpus. Vaid näitleja lubas endale vabaduse kõnes kaks lauset kokku lugeda. Väga tugev seos süntaksi ja prosoodia vahel on ka kooloni ja mõttekriipsu korral. Kaks kolmandikku koma-

dest on seotud pausidega. Kõige vähem markeeritakse kõnes nende rinnastavate sidesõnadega algavaid fraase, mis üldjuhul koma ei nõua (ja, ning, ega, ehk, või, kui ka).

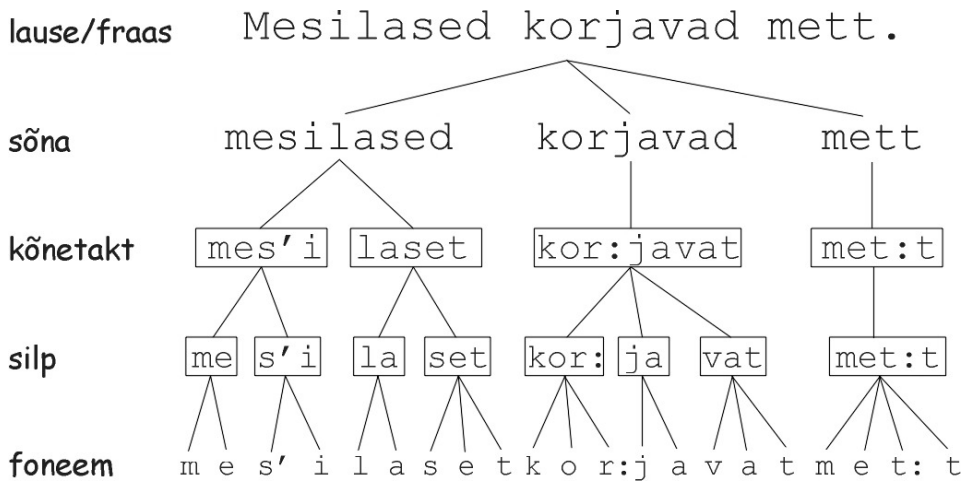
Lõpupikendusega on kirjavahemärkidest selgeim seos mõttekriipsul. Ilmselt tingib selle lugeja jaoks juba märgi kuju ise – pikk kriips kutsub esile sõnade venitamise. Pauside ja lõpupikenduste omavahelisele seotusele viitab inglise keelest pärit termin „pausieelne pikendus" (prepausal lengthening). See termin kehtib antud eestikeelse kõnematerjali põhjal vaid 60% ulatuses (601 pausist oli eelneva taktipikendusega vaid 360 pausi). Lehiste läbiviidud tajutestide põhjal (Lehiste, Fox 1993) eeldavadki eestlased lause viimasel silbil oluliselt väiksemat lõpupikendust kui näiteks inglise keele kõnelejad.

Eelnev analüüs näitas, et pausidel on kõnes väga suur variatiivsus, kuid eri liiki pausid on kestuse poolest eristatavad, pausieelsed pikendused aga mitte. Vaevalt, et sünteeskõne rütm ja loomulikkus sellest oluliselt paraneks, kui me iga teise koma järel ja iga kolmanda sidesõna ees teeksime konstantse, fraasilõpu pausi. Kõne loomulikkus pigem eeldaks, et me oskaksime pauside kestuse variatiivsust kui ka nende kõnevoos paiknemist, sünteeskõnes mõistlikult edasi anda.

6.2. Tunnuste valik segmentaalkestuste modelleerimiseks ja eksperthinnangud

Peaaegu kõigis statistilistes mudelites on faktorite ja tunnuste kestusmudelisse valikul lähtutud suuremal või vähemal määral Dennis Klatt'i reeglipõhise mudeli ideedest (Klatt 1979): kõnesegmentidel on omakestus, nad on mõjutatud naabersegmentidest, segmendi kestus sõltub tema asendist silbis, sõnas ja fraasis, aga ka üldisest kontekstist – silbi, sõna ja fraasi pikkusest. Rõhu-ajastus rütmiga keeltes on olulised ka silbi rõhulisus ja sõna esiletõsterõhk. Lisaks üldistele tunnustele sisaldavad kestusmudelid ka spetsiifilisi foneetilisi teadmisi vastava keele kohta. Näiteks on saksa keele segmentide ajalise struktuuri prognoosimudelis tunnus silbistruktuuri kohta (Möbius, van Santen 1996), aga ka hindi keeles on silbistruktuur oluline (Krishna, Taludar, Ramakrishnan 2004). Petr Horák tõi tšehhi keele kestusmudelisse ühesilbiliste sõnade eritunnuse (Horák 2005). Hollandi keeles on sarnaselt tšehhi keelega kliitikute eritunnus, aga ka sõna sageduse faktor (Klabbers 2000). Seega eeldatakse, et harjumuspäraseid, sagedamini esinevaid sõnu häältdatakse pisut erinevalt kui tekstis harva esinevaid. Keeltes, kus abisõnade hulk on küllalt kõrge, eristatakse abisõnu (e funktsioonisõnu) täistähenduslikest sõnadest (Brinckmann, Trouvain 2003; Klabbers 2000). Martti Vainio kaasas soome keele tekst-kõne sünteesi prosoodia modelleerimisel morfoloogilisi tunnuseid ja sõnaliigi infot (Vainio 2001).

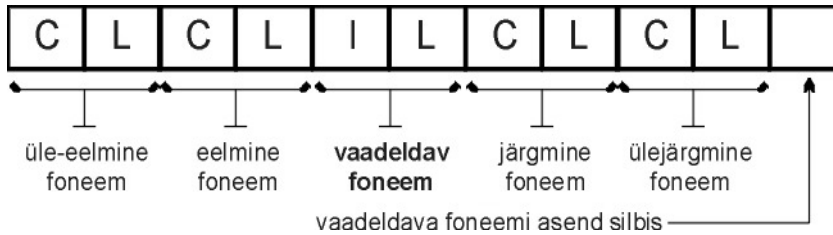
Eestikeelse kõne segmentaalkestuste modelleerimiseks lähtuti tunnuste valikul põhimõttest, et rõhu ja väldete käsitlus tugineb prosoodilisele hierarhiale, mille järgi lausung jaguneb eri tasandeis alluvussuhteis olevaiks koostisosadeks (Eek, Meister 2004:253). Joonisel 6 on näha, et lause või fraas[4] koosneb prosoodilistest sõnadest, sõnad kõnetaktidest, taktid silpidest ja kõige alumise segmentaaltasandi moodustavad foneemid. Kõigis segmentaalkestusi prognoosivais töödes ([P1], [P2], [P6], [P7] ja [P8]) esitatakse mingi kõneüksuse suhteline asukoht lauses hierarhilises mõõtkavas: foneemi asukoht silbis, silbi asukoht kõnetaktis, kõnetakti asukoht sõnas, sõna asukoht lauses. Lisaks on eelneva analüüsi põhjal osutunud oluliseks informatsioon, mis iseloomustab prosoodilise hierarhia tasandeid: silbi rõhulisus, kinnine vs lahtine silp, kõnetakti välde, fraasi pikkus sõnades jms. Paljuski põhineb selline tunnuste süsteem Klati reeglipõhisel kõne ajalise struktuuri mudeli parameetritel. Eesti keele omapäraks on kõnetakt fonoloogilise tasandina. Tšehhi uurija Pavel Horáki eeskujul (Horák 2005) on paaris viimases töös ([P6] ja [P8]) tunnuste hulka lisatud ka ühesilbilise sõna tunnus, mis osutus modelleerimisel oluliseks tunnuseks.



Joonis 6. Kõneüksuse hierarhiline kodeerimine fonoloogilises struktuuris. Näiteks foneemi [l] asend kodeeritakse vastavalt tema positsioonile kahefoneemilises silbis [la], silbi [la] asend kodeeritakse vastavalt tema positsioonile kahesilbilises taktis [laset] ja kõnetakti asend vastavalt tema positsioonile sõnas [mesilased] jne.

[4] Eesti keeles on fraasid (nimisõna-, verbi-, määrusefraas) sageli lausetes omavahel väga tihedasti põimunud, mistõttu käesolevas töös käsitletakse fraasina osalauset või loetelu elementi, mis on lausesiseselt piiritletud kirjavahemärgi või sidesõnaga. Joonisel 6 toodud näites on lause ja fraas võrdsustatud.

Järgmiseks tunnuste valiku põhimõtteks on fakt, et igal häälikul on omakestus ja et häälik on mõjutatud naaberfoneemidest. Mitu naaberfoneemi nii paremalt kui vasakult mõjutavad uuritava foneemi kestust? Esimestes töödes ([P1], [P2]) arvestati vaid ühe naaberfoneemi mõjuga nii paremalt kui vasakult suunalt. Viimastel eksperimentidel ([P6], [P8]) on osutunud optimaalseks kaasata foneemi ümbrusesse kaht naaberfoneemi (so paremalt järgmist ja ülejärgmist ja vasakult eelmist ja üle-eelmist vt joonis 7). Foneemi kirjeldab foneemiklass (9 klassi, sh ka paus) ja kontrastiivne pikkus (lühike vs pikk).



Joonis 7. Vaadeldava foneemi asukoha kodeerimine sõltuvalt ümbrusest.
(C – foneemi klass, L – foneemi kontrastiivne pikkus, I – vaadeldava foneemi identiteet).

Optimaalseks on osutunud kirjeldada foneemi ja tema ümbrust 10 tunnusega, foneemi hierarhilist asukohta lausungis kodeeritakse 5 tunnusega, osade kõneüksuste omadusi (silbirõhk, silbitüüp, kõnetakti välde) iseloomustatakse 3 tunnusega ning informatsiooni kõrgemal tasanditel olevate kõneüksuste (silp, kõnetakt, sõna, fraas, lause) pikkuste kohta 5 tunnusega. Lisaks kasutatakse binaarset tunnust, mis viitab kirjavahemärkidele etteloetavas tekstis mingi sõna järel. Kõik need tunnused (kokku 24 tunnust) moodustavad baastunnuste vektori kestusmudeli sisendisse [P8]. Algtunnuste valikul oli oluline seegi, et kõik nad oleksid sisendteksti põhjal automaatselt genereeritavad. Kõigis segmentaalkestuse modelleerimistöödes ([P1], [P2], [P6], [P7] ja [P8]) kasutati lausestajat, silbitajat, morfoloogilist analüsaatorit, ühestajat jt mooduleid, mis on loodud eesti keeletehnoloogide poolt (Viks 2000; Kaalep, Vaino 2001).

Kui tunnuste esialgne valik on tehtud, siis on võimalik saada ekspertidelt hinnang valitud argumenttunnuste vektorile ja soovitusi uute tunnuste lisamiseks. Eksperdid pidid hindama, kas mingi tunnus on nende arvates kõne ajalise struktuuri (nt segmentaalkestuste) prognoosimisel oluline või mitte, samuti küsiti nende arvamust tunnustevaheliste võimalike koosmõjude kohta. Esimestel katsetel statistilise modelleerimise valdkonnas küsisime kuuelt eesti foneetikult ja kõnetehnoloogiaga seotud inimeselt hinnanguid meie poolt valitud algsele argumenttunnuste vektorile. Ekspertide arvamused võrrelduna esimeste eksperimentide tulemustega langesid kokku vaid 41–65% ulatuses [P2]. Aga kõnematerjali lisandumisel ja eestikeelsete prosoodiliste kõnekorpuste mahu kasvades on viimaste modelleerimiseksperimentide tulemuste ja

ekspertide arvamuste kokkulangevus suurenenud [P8]. Kuid siiski küllalt suur erinevus ekspertarvamuste ja sidusast kõnest saadud tulemuste vahel on seletatav sellega, et foneetikute nn „kestusmallid" põhinevad suuresti laboratoorse kõne (isoleeritud laused ja sõnad) põhjal tehtud mõõtmistel. Isoleeritud lausete häälikukestused erinevad märgatavalt sidusa kõne temporaalsest struktuurist (Campbell 2000:312–315).

Kokkuvõtvalt, sisendteksti põhjal genereeritakse iga foneemi kohta kuni 24 tunnust, mis kirjeldavad vaadeldavat foneemi ennast ja tema ümbrust, paiknemist hierarhilises süsteemis ja kõrgemate tasandite üksuste omadusi. Tunnuste valikul ja nende omavaheliste seoste määratlemisel tasub nõu küsida ekspertidelt.

6.3. Statistiliste meetodite võrdlus kestuste prognoosimisel

Milline on hea meetod kõneprosoodia prognoosimisel? Kas on olemas objektiivseid kriteeriume parima statistilise meetodi valikuks? Nende küsimustega puutub kokku iga uurija, kes püüab statistiliste meetoditega sidusa kõne põhjal kõneprosoodiat modelleerida. Esimestes modelleerimiseksperimentides ([P1], [P2]) kasutati põhiliselt mitmest lineaarset regressiooni. Peaaegu alati tekib kahtlus, kas minu valitud meetod on ikkagi küllalt hea või eksisteerib veelgi parem. Autori jaoks oli esimeseks tõukeks nende küsimuste üle juurdlemiseks Yoshinori Sagisaka plenaarettekanne foneetikateaduste kongressil Barcelonas 2003. aastal, kus ettekandja väitis, et neil on üle kahekümne aastane kogemus kõneprosoodia modelleerimise vallas ja nemad eelistavad regressioonanalüüsi meetodeid (Sagisaka 2003). Tutvudes erinevate töödega antud vallas (Brinkmann, Trouvain 2003; Horak 2005; Krishna, Murthy 2004; Vainio 2001), torkab silma, et regressioonanalüüsi meetodite asemel kasutatakse valdavalt närvivõrke ja regressioonipuid kõneprosoodia prognoosimiseks. Konkreetse masinõppe meetodi valikut tavaliselt ei põhjendata ning saadud prognoosi tulemusi võrreldakse enamasti olemasoleva reeglipõhise prosoodia generaatoriga. Tundub, et konkreetse meetodi valik on pragmaatiline, sõltudes uurija enda hariduslikust taustast, juhendajate ja kolleegide mõjutustest, vastava tarkvara kättesaadavusest ja muudest põhjustest.

Statistikaprogrammipaketi SAS 9.1 litsentsi omandamisega tekkis hea võimalus võrrelda erinevaid prognoositehnikaid (regressioon, CART meetod, närvivõrgud) omavahel ühel ja samal andmehulgal häälikute segmentaalkestuste prognoosimisel [P6]. Meetodeid hinnati prognoosivea, tulemuste interpreteeritavuse, andmete eeltöötluse vajaduse jm kriteeriumite alusel.

Algandmeteks olid mees- ja naisraadiodiktorite kõneandmed. Teksti põhjal genereeriti 26 argumenttunnust. Tunnuste hulga optimeerimiseks tehti andmete põhjal eelanalüüs. Lineaarse regressioonanalüüsiga selgitati välja need tunnused,

mis osutusid oluliseks nii mees- kui ka naisdiktori andmehulgal genereeritud prognoosimudelis. Kokku oli selliseid tunnuseid 18 (vt [P6] tabel 1).

Mudelite funktsioonitunnuseks kõigi kolme meetodi korral oli häälikute logaritmitud kestused. Ehkki närvivõrgud ja CART-meetod otseselt ei nõua funktsioonitunnuse normaaljaotust, tuleb närvivõrkude stabiilsusele normaliseerimine kindlasti kasuks.

Statistilise modelleerimise tulemused on toodud tabelis 2. Kõik võrreldavad meetodid andsid samal andmekogumil ja samade argumenttunnuste juures väga sarnase veaprotsendi. Eriti üllatas just see, et lineaarsel regressioonil saadi peaaegu kõige madalam veaprotsent. Oma olemuselt peaks ju lineaarne regressioon tuvastama vaid kõige otsesemaid ja ilmsemaid seoseid sisendi ja väljundi vahel. Ja ehkki funktsioonitunnuse logaritmimisega ning sisendtunnuste mittelineaarse kodeerimisega on regressioonimudelisse salvestunud ka teatud hulk mittelineaarsust, peaks siiski varjatumad seosed sisendi ja väljundi vahel tulema ilmsiks keerukamate mittelineaarsete meetodite (so klassifikatsioon ja regressioonipuud ning närvivõrgud) abil. Seega lineaarse regressiooni meetod, mida on kasutatud kaua ja edukalt kõnelaine töötluseks (Markel, Gray 1976) ning mis leiab siiamaani rakendust kõne analüüsil ja sünteesil, on arvestatav meetod ka kõne ajalise struktuuri modelleerimiseks.

Tabel 2. Prognoosivead ja teised mudelite hindamiskriteeriumid.

Kriteeriumid		*Närvivõrgud*	*Regressioon*	*CART*
Prognoosivead:	Treenimisel	0,230	0,230	0,250
– meesdiktor	Valideerimisel	0,243	0,248	0,264
(keskmine viga 21%)	Testimisel	0,230	0,232	0,255
	Treenimisel	0,224	0,221	0,230
– naisdiktor	Valideerimisel	0,221	0,218	0,231
(keskmine viga 19%)	Testimisel	0,221	0,217	0,230
Mudeli interpreteerimine		keeruline	lihtne	väga lihtne
Väljundi normaliseerimine		soovitav	vajalik	mittevajalik
Sisendite eeltöötlus		vajalik	vajalik	mittevajalik
Interaktiivne treenimine		ja	ei	ja
Mudel puuduvate sisendväärtustega		ei	ei	ja

Kõige selgemini on mudel interpreteeritav kahendpuul, üsna selgesti on sisendi mõju kestusele arusaadav regressioonikoefitsientide analüüsil. Hoopis raskem on õpiprotsessi tulemusi tõlgendada närvivõrkudel. Lineaarne regressioon eeldab funktsioonitunnuse normaaljaotust, teised meetodid seda otseselt ei nõua, aga närvivõrkude mudeli stabiilsust normaliseerimine parandab. Enne statistilist modelleerimist regressioonanalüüsil ja närvivõrkudel tuleb argumenttunnuseid töödelda. Regressioonianalüüsi tarvis on vaja nominaalsete tunnuste asemele genereerida suur hulk binaarseid pseudotunnuseid. Närvivõrkude sisendite

muutumispiirkond tuleb teisendada vahemikku [0, 1]. Tabelis 2 toodud kaks viimast tunnust on kaudsemad kriteeriumid meetodite üle otsustamiseks.

Seega lineaarne regressioon on prognoositäpsuselt konkurentsivõimeline keerukamate mittelineaarsete meetoditega (CART-meetod, närvivõrgud). Mudeleid on kõige parem tõlgendada regressioonipuudel.

6.4. Leksikaalne prosoodia

Traditsiooniliselt ei ole nende faktorite hulgas, mis mõjutavad märkimisväärselt kõne ajalist struktuuri sõnaliigi infot ja morfoloogilisi tunnuseid (van Santen 1998, Campbell 2000, Sagisaka 2003). See võib olla seotud sellega, et enamik uurimustest tekst-kõne sünteesi kohta on kontsentreerunud keeltele, kus morfoloogial on võrdlemisi väike roll. Üheks keeleks, mille kohta on hinnatud morfoloogiliste tunnuste osatähtsust kõneüksuste kestusele, on soome keel (Vainio 2001). Eesti keeles on sõnal väga tähtis roll nii grammatikas kui ka foneetikas ja väga rikas morfoloogia. Seetõttu tekkis huvi kontrollida, kas morfoloogilised, leksikaalsed ja võib-olla ka süntaktilised tunnused mõjutavad kõne ajalist struktuuri eesti keeles [P7]. Ilmselt kõige loomulikumaks viisiks morfoloogia, sõnaliigi ja süntaksi faktorite mõju arvestamiseks oli laiendada meie varasemat statistilise modelleerimise metodoloogiat ja uurida, kuidas nad mõjutavad kestusmudelite toimimist. Modelleerimiseks valiti kaks erinevat meetodit: lineaarne regressioon ja mittelineaarne närvivõrkude meetod. Faktorite mõju kvalitatiivseks hindamiseks mõõdeti ja võrreldi mudelite väljundvea muutust. Tehtud eksperimendid näitasid väljundvea mõneprotsendilist vähenemist, kui kestusmudeli sisendisse oli lisatud morfoloogilis-süntaktilist ja sõnaliigi infot.

Saadud kestusmudelid [P7]-s põhinesid kahe raadiodiktori kõnel, mistõttu oli mudelite interpreteerimisel üldistusi veel vara teha. Aga kõige selgemad seaduspärasused ilmnesid sõnaliigi faktori parameetrite regressioonikordajate visuaalsel ülevaatusel regressioonmudelis. Tabelis 3 on toodud häälikute keskmised lühenemised-pikenemised sõnaliigiti verbihäälikute suhtes mees- ja naisdiktori kestusmudelis. Tabeli keskosas on varieeruvus suurem, aga tabeli algus ja lõpp on väga sarnased. Tabeli 3 põhjal võib öelda, et pärisnimede hääldamisel on häälikud sõnas keskmiselt 5–6 millisekundit pikemad. Keskmine häälikute pikkus oli neil diktoritel vastavalt 62,5 ja 64,1 millisekundit. Seega hääldasid nad pärisnimesid ca 10% pikemalt kui verbe. Natuke pikemalt hääldati nimisõnu ja kaassõnu. Üllatav, et kaassõna häälikud olid keskmiselt pikemad. Kaassõna kuulub nn abisõnade klassi. Seda tüüpi sõnad on enamikus keeltes lühemad kui täistähenduslikud sõnad. Kaassõna eesti keeles kuulub alati kokku nimisõnaga, mis on tihti lauses fookuses ja keskmisest pikem ning mille mõju võib ulatuda ka tema naabruses oleva kaassõnani. Üle 10% lühemalt hääldati

jällegi järgarvsõnu ja ca 5% lühemalt asesõnu ja määrsõnu. Järgarvsõnade lühenemine on loogiliselt seletatav sellega, et tekstides on küllalt palju aastaarve, mis on enamasti väljendatud järgarvsõnana. Eelmise sajandi küllalt pikkade aastaarvude väljaütlemisel kipuvad diktorid teksti lugedes kiirustama, sest konteksti põhjal on aastaarvust tavaliselt olulised vaid viimane või kaks viimast numbrit. Aga teksti korrektsel lugemisel peab välja ütlema kogu aastaarvu.

Tabel 3. Keskmised häälikute pikenemised-lühenemised (millisekundites) sõnaliigiti.

Sõnaliik	*Meesdiktor*	*Naisdiktor*
pärisnimi	6,23	5,22
nimisõna	2,25	2,10
kaassõna	0,82	2,82
genitiivatribuut	0,42	1,35
verb	0,00	0,00
põhiarvsõna	–0,10	0,42
sidesõna	–0,14	1,81
omadussõna	–0,39	1,14
määrsõna	–0,89	–2,90
asesõna	–4,13	–3,86
järgarvsõna	–5,44	–7,48

Tulemused näitasid prognoosivea vähenemist mõne protsendi võrra, kui kestusmudeli sisendisse lisada morfoloogilis-süntaktilist ja sõnaliigi informatsiooni, mida võiski eeldada, arvestades sõna tähtsat rolli nii eesti keele grammatikas kui ka foneetikas.

6.5. Modelleerimistulemused, olulised tunnused, prognoosivead ja tulemuste interpreteerimine

6.5.1. Pauside modelleerimine

Pauside kestuste modelleerimiseks genereeriti teksti põhjal hulk tunnuseid [P3], [P5], mis kirjeldasid: teksti struktuuri (lõigu-, lause- ja fraasilõpp, sidesõnad tekstis); pausile eelnevat kõnetakti (takti pikkus häälikutes, taktivälde, takti viimase silbi pikkus häälikutes ja binaarne tunnus, mis näitas lõpupikendust); pausi ajalisi suhteid (pausi kaugus lõigu, lause ja fraasi algusest ning samuti kaugus eelnevast pausist ning eelnevast sissehingamisest).

Prognoositavaks tunnuseks oli pausi kestus. Lineaarse regressiooni tarvis tuli funktsioonitunnus logaritmida, kuna logaritmitud kestus allub enam normaaljaotusele.

Pauside kestuste modelleerimisel mitmesel regressioonil osutusid olulisteks tekstistruktuuri tunnustest lõigu-, lause ja fraasilõpp. Pausile eelneva kõnetakti tunnustest osutus oluliseks usaldusnivool 0,05 vaid binaarne tunnus, mis näitas, kas pausile eelnev kõnetakt oli pikendatud või mitte. Lisaks oli kestuse prognoosimisel oluline konkreetse pausi kaugus talle eelnevast pausist. Joonisel 8 on toodud pausi kestuse arvutusvalem logaritmilisel kujul.

LN(pausi kestus) *= –1,973 + 0,373 * LQLQP + 1,454 * LALQP + 0,441 * FRKOM + 0,012*KAUGFR+0,024 * KAUGPA + 0,133 *PIKENDUS*

Joonis 8. Regressioonvõrrand pausi logaritmilise kestuse arvutamiseks. Muutujad: LQLQP – lõigulõpu tunnus, LALQP – lauselõpu tunnus, FRKOM – fraasilõpp (koma), KAUGFR – eelneva fraasi pikkus, KAUGPA – kaugus eelmisest pausist, PIKENDUS – viimase kõnetakti pikendus.

Pauside asukoha prognoosimiseks rakendati logistilist regressiooni, millega prognoositi tõenäosust, kas antud sõna järel kõnevoos tehakse paus või mitte. Logistilise regressiooni muutujatena kasutati suures osas samu tunnuseid, mis pauside kestuste ennustamisel. Aga oli ka mõningaid erinevusi.

Artikli [P4] kõnematerjali analüüsil täheldati, et pauside asukoht kipub korreleeruma lisaks kirjavahemärkide ja sidesõnadega ka pärisnimede ja järgarvsõnadega. Hilisem statistiline analüüs mahukamal kõnematerjalil tunnistas need korrelatsioonid väheolulisteks. Ka artiklis [P3] kaasati sisenditena kaks binaarset tunnust, mis näitasid, kas järgnev sõna on pärisnimi või võõrsõna. Neid tunnuseid ärgitas lisama kujutelm, et pärisnimede ees (nt *Minu nimi on Tamm, Jüri Tamm.*) ja võib-olla ka enne keerulisemate võõrsõnade väljaütlemist (nt *Rahvas toetas konstitutsioonilist monarhiat.*) tehakse kõnes väikene paus. Aga ka see hüpotees ei leidnud tõestust. Pärisnimede ja võõrsõnade korrelatsioon pausiga oli väga nõrk ja need tunnused osutusid ebaolulisteks [P3].

Tabel 4. Logistilise regressiooni tulemused: lausesisese pausi asukohta oluliselt mõjutavad muutujad, nende šansside suhe ja usalduspiirid.

Sõltumatud muutujad	*Šansside suhe*	*Usalduspiirid*	
		alumine	*ülemine*
Sõna järel on koma tekstis	17,4	11,7	25,9
Järgmine sõna on sidesõna	7,9	4,8	12,8
Sõna kaugus lause algusest	1,1	1,0	1,2
Viimase kõnetakti pikkus	1,3	1,1	1,4
Viimase kõnetakti välde	1,2	1,1	1,5
Viimane kõnetakt on pikendatud	6,9	5,2	9,2

Tabelis 4 on toodud logistilise regressiooni poolt leitud kuus tunnust, mis mõjutavad lausesisese pausi asukohta. Väga oluline tegur lausesiseseks pausiks on koma tekstis, šanss pausiks on sel juhul 17,4 korda keskmisest suurem. 7–8 korda suurendab šanssi sõnajärgseks pausiks see, kui järgnev sõna on sidesõna või kui selle sõna kõnetakt on pikendatud. Keskmisest pisut sagedamini tehakse kõnes paus pikemate kõnetaktide ja kõrgemavälteliste sõnade järel. Nende tunnuste osa kipub siiski prognoosimudelis marginaalseks jääma, kuna nad tõstavad šanssi pausi esinemiseks vaid 1,2–1,3 korda.

Kokkuvõtteks võib öelda, et eri liiki pausid on kõnes kestuselt eristatavad. Pauside kestus ja nende asukoht kõnevoos on modelleeritav statistiliste meetoditega. Johtuvalt pauside suurest variatiivsusest, kirjeldavad saadud mudelid diktoritevahelisi keskmistatud hääleparameetreid ja kõige üldisemaid seaduspärasusi pauside kestustes ja nende paiknemises. Seega peaks paari diktori kohta koguma mahuka kõnematerjali ja sama metoodikat rakendama ühe diktori andmestikul eraldi. Lõpupikenduste prognoosimiseks ei õnnestunud arvestatavat kestusmudelit luua. Ilmselt tuleb lõpupikendusi käsitleda häälikute kestusmudeli osana.

6.5.2. Segmentaalkestuste modelleerimine

Kestuste mudelis on argumenttunnuste hulgas arvukalt nominaalseid muutujaid: vaadeldava foneemi identiteet (26 foneemi), naaberfoneemide klassid (8 foneemiklassi + paus), lisaks veel võimalikud nominaalsed morfoloogilised, süntaktilised ja sõnaliigi tunnused (vt tabel 5). Seega võib regressioonvõrrandi parameetrite hulk ulatuda kuni sajani ning kestusmudelit on mõistlik tõlgendada argumenttunnuste olulisuse seisukohast. Vaid esimestes töödes ([P1] ja [P2]), kus sisendite hulk oli väiksem ja kestust prognoositi vaadeldava foneemi klassi tasandil, osutus võimalikuks tulemusi ka võrrandina esitada. Kõige selgemalt on tunnuste tähtsus määratav regressioonmudelis, kus igale tunnusele esitatakse statistiline hinnang tema olulisusele. Näiteks *forward selection*-meetodi korral hakatakse ühe kaupa mudelisse lisama antud hetkel kõige olulisemaid tunnuseid. Ümberhinnang toimub iga tsükli eel. Ka CART-meetodi korral satuvad regressioonipuusse kõige olulisemad tunnused. Närvivõrkude meetodis selline olulisuse hinnang puudub ning seal saab otsustada mingi tunnuse vajalikkuse üle, tunnuseid käsitsi lisades või eemaldades ning väljundit hinnates. Esimestel modelleerimiseksperimentidel saadud tulemused tunnuste olulisuse osas näisid „tähtsate avastustena" kõneprosoodia vallas, nt välte klassifitseerimine eba-oluliseks tunnuseks [P1], [P2]. Hilisemad eksperimendid ([P6], [P7], [P8]) osutasid, et oluliste tunnuste hulk võib diktoriti kõikuda, ka ei pruugi erinevatel meetoditel tähtsad tunnused alati üks-ühele kattuda.

Tabel 5. Sisendite e argumenttunnuste olulisus häälikute segmentaalsete kestuste modelleerimisel

	Sisendid
1.	**üle-eelmise foneemi klass**
2.	**üle-eelmise foneemi pikkus**
3.	**eelmise foneemi klass**
4.	eelmise foneemi pikkus
5.	**vaadeldava foneemi identiteet**
6.	**vaadeldava foneemi pikkus**
7.	**järgmise foneemi klass**
8.	järgmise foneemi pikkus
9.	**ülejärgmise foneemi klass**
10.	**ülejärgmise foneemi pikkus**
11.	**foneemi asend silbis**
12.	silbi rõhulisus
13.	silbi tüüp
14.	kõnetakti välde
15.	**silbi asend kõnetaktis**
16.	**kõnetakti pikkus silpides**
17.	kõnetakti asend sõnas
18.	**sõna pikkus taktides**
19.	**ühesilbiline sõna**
20.	**sõna asend fraasis**
21.	**fraasi pikkus sõnades**
22.	lause pikkus fraasides
23.	**kirjavahemärgid**
24.	**morfoloogia**
25.	**sõnaliik**
26.	**süntaks**

Tabelis 5 on arvukate eksperimentide põhjal määratud tunnuste olulisused häälikute kestuste prognoosimisel [P8]. Poolpaksus kirjas tumedamal foonil on toodud need tunnused, mis on olulised enamiku (üle 80 %) diktorite kohta. Tavalises kirjas heledamal hallil taustal on need tunnused, mis on osutunud mõnede (alla poolte) diktorite kestusmudelites ebaolulisteks. Huvitaval kombel kõnetakti välde, mis on eesti keele sõnaprosoodia nurgakiviks, on osutunud mõne diktori andmetel kestuste prognoosimisel ebaoluliseks tunnuseks. Siin võib olla põhjuseks see, et väldet kui suprasegmentaalset nähtust ei saa esitada ühe lineaarse tunnusena. Välteid kui vastanduvaid kvaliteedimalle seostatakse kõnelõiguga, mis algab kõnetakti rõhulise silbi vokaalist kuni rõhuta silbi vokaali lõpuni (Eek, Meister 2004). Seega on kõnes suur hulk häälikuid (rõhulise silbi alguses (onset), rõhuta silbi koodas, jm), mis ei osale vältevastanduses. Selliste kõneüksuste kestuste prognoosimisel on ilmselt kõnetaktivälde ebaoluline tunnus. Ilmselt peaks hääliku asendit kirjeldavasse hierarhilisse süsteemi

kaasama silbistruktuuri iseloomustava tasandi. Silbistruktuur kajastub saksa (Möbius, van Santen 1996) ja hindi keele (Krishna jt 2004) kestusmudelis. Võimalikele koosmõjudele rõhu ja silbistruktuuri vahel viitab see, et ka silbirõhk ei mõjuta alati oluliselt prognoositavat kestust. Üllatavalt on järgneva hääliku kontrastiivne pikkus vähem oluline kui ülejärgmise hääliku pikkus [P8]. Tavalises kirjas valgel taustal on kaks tunnust, mis osutusid süstemaatiliselt ebaolulisteks segmentaalsete kestuste prognoosimisel: silbi tüüp (lahtine vs kinnine silp) ja kõnetakti asend sõnas.

6.5.3. Mudelite olulisus ja prognoositäpsus

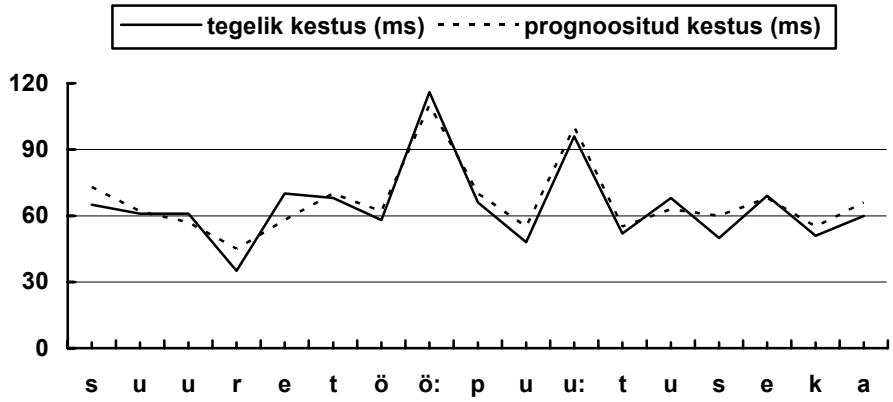
Modelleerimise edukuse üle saab otsustada kõigepealt selle järgi, kas genereeritud mudel on statistiliselt oluline, ja teiseks selle põhjal, kui suure osa funktsioonitunnuse varieeruvusest mudel suudab kirjeldada. Otsustamisel peab kontrollima kogu mudeli statistilist tähtsust. Võib kokkuvõtvalt öelda, et kõik väitekirja eksperimentides genereeritud mudelid on statistiliselt olulised. Tabelis 6 on näitena toodud pauside kestuse regressioonmudeli summaarne sobivus ja dispersioonanalüüs. Näeme, et mudel on statistiliselt oluline, lineaarne seos 9 sisendi ja pausi kestuse vahel on usaldatav (mudeli olulisuse tõenäosus on peaaegu null). Seos on üsna tugev (korrelatsioonikordaja r =0.83), mis kirjeldab 2/3 pausi kestuse varieeruvusest (determinatsioonikordaja r^2=0,67). Erinevad pauside kestusmudelid kirjeldasid 65–73% kestuste variatiivsusest, häälikute kestusmudelis jäi vastav näitaja 52–63% piiridesse [P6], [P7], [P8], mis samuti on hea näitaja. Lõpupikenduste modelleerimine ebaõnnestus, kuna saadud mudelid kirjeldasid vaid 25–30% pikenduste muutumisest [P5].

Tabel 6. Pauside kestuste regressioonmudeli summaarne sobivus ja dispersioonanalüüs.

Summaarne sobivus					
korrelatsioonikordaja 0,82673		determinatsioonikordaja 0.6686			
Dispersioonanalüüs					
varieeruvuse allikas	vabadus-astmete arv	hälvete ruutude summa	keskruut	F-statistik	mudeli olulisuse tõenäosus
mudel	9	478.5	18.403	220.87	<0.0001
prognoosijäägid	560	408.8	0.0862		
Kokku	559	787.2			

Pauside asukoha mudeli prognoositäpsus küündis 88 protsendini, st hinnang sellele, kas mingi sõna järel on paus või mitte. Analüüsi põhjal leidsime, et osa tekstistruktuuri märgendite (lõigulõpp, lauselõpp, koolon, mõttekriips) järel on paus 93–100% tõenäosusega. Kui sellised „kindlad" pausid mudelist välja jätta, siis saadud lausesiseste pauside mudel suudab prognoosida lausesiseseid pause vaid 44% täpsusega [P5].

Häälikute kestuste prognoosiviga erinevates eksperimentides jääb piiridesse 16,1–21,2%, sõltudes diktorist ja meetodist. Testides eri meetodeid (lineaarne regressioon, CART ja närvivõrgud) samal andmehulgal, selgus, et lineaarne regressioon ja närvivõrgud olid prognoosimistäpsuselt peaaegu võrdsed, CART mudeli prognoosiviga oli pisut suurem [P6]. Oli üllatav, et lineaarne meetod suutis konkurentsi pakkuda mittelineaarsetele meetoditele, ehkki lineaarne mudel peaks näitama vaid kõige ilmsemaid ja üldisemaid seoseid sisendite ja väljundi vahel ning vaid mittelineaarsete meetoditega võiksid esile tulla varjatumad seosed. Joonise 9 graafikul on toodud kõnelõigu häälikute tegelik kestus vs närvivõrkudel prognoositud kestus (Fišel, Mihkla 2006). [P5]-s toodud üldises pauside kestusmudelis on veaprotsendiks ekslikult märgitud 29–37%, tegelik viga on poole väiksem so 14–18%. Ühe diktori kõnematerjali põhjal genereeritud mudeli prognoosiviga jääb sõltuvalt diktorist 8–12% piiridesse.



Joonis 9. Tegelik häälikute kestus versus prognoositud kestus närvivõrkude mudeliga.

Kokkuvõtvalt võib modelleerimistulemuste kohta öelda, et pauside kestusmudel näitab väga tugevat seost väljundi ja argumenttunnuste vahel. Pauside asukohad on tugevasti korreleerunud teksti liigendusega: kirjavahemärkide ja sidesõnadega. Piiripikenduste modelleerimiseks ei õnnestu usaldusväärset mudelit luua, samuti on fraasisiseste pauside asukohta keeruline määrata.

Segmentaalkestuste prognoosimudel sisaldab küllalt suure hulga faktoreid ja tunnuseid. Vaadeldava foneemi ümbrust on optimaalne kirjeldada mõlemalt poolt kahe naaberfoneemiga. Olulised tunnused on foneemi asend lausungi hierarhilises struktuuris ja kõrgemate fonoloogiliste tasandite (silp, takt, sõna, fraas) omadused, teksti liigendus ning ka morfoloogilis-süntaktiline ja sõnaliigi info.

Kõik mudelid on statistiliselt olulised ja kirjeldavad 65–73% pauside ja 52–63% segmentaalfoneemide kestuste variatiivsusest.

7. KOKKUVÕTE JA EDASISE TÖÖ SUUNAD

See väitekiri on vaid pika tee alguseks kõiki faktoreid haarava eestikeelse kõne ajalise struktuuri juhtimismudeli loomisel. Selle töö põhiline panus seisneb erinevatel statistilistel tehnikatel põhineva metodoloogia väljatöötamises kõnekorpuste baasil eesti keele prosoodia modelleerimiseks ja uurimiseks.

Arvukatel modelleerimiseksperimentidel ja statistilisel analüüsil saadud tulemustest võiks esile tuua järgmist:

- töö käigus koguti 27 keelejuhi loetud tekstide sidusa kõne korpus;
- loetud tekstides on pausid kestuselt klassifitseeritavad lõigu-, lause- ja fraasilõpu pausideks;
- pauside kestused ja nende asukohad on kõnevoos prognoositavad, kusjuures kõige tugevam korrelatsioon on teksti liigendusega (kirjavahemärgid, sidesõnad), aga ka kaugusega eelmisest pausist ja lause algusest;
- segmentaalkestuste prognoosimisel osutusid olulisteks tunnused, mis kirjeldasid hääliku naaberfoneemide mõjusid mõlemalt poolt kaheses aknas (eelmine ja järgmine ning üleeelmine ja ülejärgmine foneem), aga samuti hääliku hierarhilist paiknemist lausungi fonoloogilises struktuuris (foneemi asend silbis, silbi asend taktis, sõna asend fraasis jms); lisaks veel tunnused, mis iseloomustasid foneemi klassi, silbi rõhulisust, sõna ühesilbilisust, fraasi pikkus sõnades, jne;
- häälikukestuste modelleerimisel mängis otsustavat rolli veel teksti liigendus (kirjavahemärgid ja sidesõnad);
- sõnade süntaktilised, morfoloogilised ja sõnaliigi tunnused mõjutavad sõna moodustavate segmentide kestusi, kõige paremini interpreteeritavad tulemused ilmnesid sõnaliigiti;
- erinevate meetodite võrdlemisel ilmnes, et lineaarne regressioon on võrdväärne statistiline meetod prognoositäpsuselt kestuste ennustamisel CART-meetodi ja närvivõrkudega; tulemuste interpreteeritavus on parim CART-meetodi korral, mille rakendamine nõuab aga foneetiliselt tasakaalustatud kõnekorpust.

Kõne ajalise struktuuri modelleerimiskogemused lubavad väita, et kõnekorpustel põhinevad statistilised tehnikad võimaldavad usaldusväärselt prognoosida segmentaalkestusi ja vältida suuri vigu, mis võivad olla põhjustatud juhtimisreeglite halvast kombinatsioonist. Lisaks on statistiliste meetoditega võimalik avastada ja uurida väikesi, kuid olulisi erinevusi ajalises struktuuris, näiteks häälikute kestuste sõltuvus sõnaliigist [P7].

Kõne ajalise struktuuri häälikukestuste ja pauside täpsem modelleerimine tekst-kõne sünteesi tarbeks parandab sünteeskõne kvaliteeti ja annab võimaluse kõnekorpuste baasil automaatselt genereerida sünteesiks erinevaid hääleprofiile.

Mida oleks võinud töös teisiti teha? Materjali valikul oleks võinud piirduda vähema arvu diktoritega ja koguda mahukam kõnematerjal mõne diktori kohta,

samuti oleks võib-olla tulnud piirduda ühe konkreetse tekstitüübiga (nt uudised). BABELi foneetilises andmebaasis olid ühe diktori loetud kõnelõigud suhteliselt lühikesed, mis ei võimaldanud luua pauside mudelit konkreetse diktori kohta ja mõne meetodi korral kippus ka diktori segmentaalkestuste mudeli jaoks materjali nappima. Teise võimaliku täpsustusena peab mainima, et tunnuste hulka oleks tulnud lisada infot silbistruktuuri kohta, sest kõnetakti välde avaldub kõige selgemini just lõigus rõhulise silbi vokaalist kuni rõhuta silbi vokaali lõpuni (Eek, Meister 1997; Ross, Lehiste 2001). Sellele juhitakse tähelepanu ka punktis 6.5.2, et mitte kõik takti moodustavad häälikud pole kõnetakti välte identifitseerimisel samaväärselt olulised.

Edasiste tööde kavandamisel peaks arvestama nende potentsiaalsete võimalustega. 2006. a. sügisel käivitus eestikeelse korpuspõhise sünteesi projekt (Mihkla jt 2007) riikliku programmi „Eesti keele keeletehnoloogiline tugi" raames. Korpuspõhise projekti kõneandmebaasid sisaldavad juba ca 50 minutit kõnet ühe diktori kohta. Kõnekorpuse aluseks on foneetiliselt „rikkad" tekstid, mis sisaldavad kõiki difoone, sagedasi sõnu ja fraase, palju sõnavorme, numbreid ja aastaarve (Piits jt 2007). Need kõnekorpused on heaks baasiks töös väljapakutud metodoloogia rakendamiseks kõne ajalise struktuuri modelleerimiseks.

Artiklites [P6], [P7] on ka viidatud tajukatsete korraldamise vajadusele. Et tekst-kõne süntesaatori põhilised kasutajad on pimedad ja nägemispuudega inimesed, siis vastavad katsed käivad koostöös Põhja-Eesti Pimedate Ühingu liikmetega.

Teiseks oluliseks uute tööde kavandamise suunaks on kõne prosoodia teiste külgede – põhitooni ja intensiivsuse – statistiline modelleerimine kõnekorpuste baasil. Põhitooni modelleerimise üksikuid külgi on põgusalt käsitletud ka väitekirja artiklites: *kas*-küsimuse intonatsiooni modelleerimine [P1] ja intonatsiooni seos süntaktiliste, morfoloogiliste ja sõnaliigi tunnustega [P4]. Samuti on ühe raadiodiktori kõnemeloodiat modelleeritud kestustunnuseid rakendades. Kuna põhitoon ja kõnesignaali intensiivsus sõltuvad teataval määral erinevatest tunnustest kui kestus, siis peaks oluliste tunnuste valikuks korraldama modelleerimiseksperimente.

SUMMARY

In the present dissertation, a methodology is presented for an automatic generation of models of the temporal structure of speech for the purposes of a high-quality Estonian text-to-speech (TTS) synthesis. The main problems of prosody modelling have always been connected with the so called "fuzzy area" between the discrete symbolic representation of speech and the continuous speech wave. An ordinary written text contains no other symbols but punctuation marks to direct the temporal structure of speech (the duration of speech units and pauses, the position of pauses, speech rate etc.). The naturalness of the temporal structure in synthetic speech, however, requires that the durations of segments and pauses as well as the position of pauses in the speech flow should not differ significantly from their values in natural connected speech. A shortcoming of rule-based prosody models in TTS synthesis is the considerable dependence of the rules on measurements made on the basis of so-called laboratory speech, and also that these models contain mistakes due to the simultaneous implementation of independently derived rules. The use of connected speech corpora and statistical optimization, however, make it possible to replace rule-writing with statistical modelling and to improve the quality of synthetic speech.

In the present study, various statistical methods (linear and logistic regression, classification and regression trees (CART), and neural networks) were applied on the corpora of connected speech in order to predict the durations of speech sounds and pauses. As the aim of the work was the modelling of the temporal structure of speech for TTS synthesis, the corpus of connected speech consisted of different types of read text (fiction, news, samples from the Estonian Phonetic Database) recorded by 27 speakers.

The modelling experiments showed that it is possible to predict the durations as well as the positions of pauses in the speech flow. The models had the strongest correlation with the structure of the text (punctuation marks and conjunctions), as well as the distance from the previous pause, and the position in the clause. Pauses in read texts can be classified automatically, on the basis of their duration, into paragraph-, clause-, and phrase-final pauses.

In predicting segmental durations, features which proved to be significant included those which describe the dependence of a given phoneme on its adjacent phonemes, and also these features which characterise the position of a phoneme in the hierarchical structure of the utterance (e.g. the position of a phoneme in the syllable, the position of a word in the phrase, etc). Additionally, statistically significant were such features which characterise the phoneme class, syllable stress, monosyllabicity etc., and the syntactic structure of the text.

In Estonian, the word and its form have a vital role both in grammar and in phonetics. The present work showed that the duration of the segments in a word is influenced by the syntactic, morphological and part-of-speech features of the word.

A comparison of different methods of prediction revealed that as far as the predictive precision is concerned, linear regression is an equally efficient statistical method as nonlinear methods (CART and neural networks).

Besides speech technology, a corpus-based modelling of the temporal structure of speech is also of interest for phonetics, as it enables, for instance, to analyze small hidden, yet important differences in segmental durations, which are caused by the morpho-syntactic structure and part-of-speech. In phonetic sciences, such corpus-based statistical methodology makes it possible to test different theoretical approaches on large amounts of data and to carry out precision analysis of numerous phenomena, thus providing a statistically grounded understanding about the operation of cognitive mechanisms in phonetics.

ACKNOWLEDGEMENTS

Work on the present dissertation was carried out in 2004–2007 at the Institute of the Estonian Language and from 2005 also in the framework of the doctoral school "Linguistics and Language Technology". Many people have put their hearts and minds into it.

First I would like to thank my supervisors Dr Einar Meister and Prof Haldur Õim for their valuable advice in relation to my research topic and also to my studies at the doctoral school. Apart from being my supervisor, Einar Meister is also a partner in a long-term fruitful cooperation in Estonian text-to-speech synthesis research. My special thanks go to Dr Arvo Eek for his valuable comments on several articles and on the summary of my dissertation. Arvo Eek was of great help in defining and describing the terms and concepts used in my dissertation.

I would also like to express my thanks to Hille Pajupuu and Krista Kerge, co-authors of articles on sentence intonation, and Jüri Kuusik who led me to statistical prediction models.

The Director of the Institute of the Estonian Language Prof Urmas Sutrop inspired me both in my doctoral studies and in writing an essential article [P8] for my dissertation for the journal *Trames*. The head of the doctoral school Prof Karl Pajusalu gave me extremely useful advice on how to write a summary of my dissertation.

I also collaborated closely with the North-Estonian Association for the Blind. The blind and visually impaired are daily users of the Estonian TTS synthesiser. They are also the best testers of the results of speech prosody modelling. Artur Räpp and Eduard Borissenko have given me constructive feedback on the functioning of speech temporal structure models in the form of comments and recommendations. Thank you!

Many thanks go to Sirje Ainsaar for her high-quality translations of articles into English, to Jana Tiitus from Tallinn University for her speedy and professional translation of the summary and to Eva Liina Asu-Garcia from Tartu University for proofreading the English of the summary and introduction. I am also grateful to my colleagues Liisi Piits and Indrek Kiissel for taking the time to read my dissertation with a critical eye and prepare it for publication.

My heartfelt thanks go to my family: my wife Külli and daughters Triin, Maarja, Laura and Liisa have been very supportive during my late studies.

Last but not least, I would like to thank all my colleagues and everybody else who have contributed to this dissertation.

Tallinn, December 2007
Meelis Mihkla

LIST OF PUBLICATIONS

The present dissertation consists of the following list of publications referred to in the text as [P1]…[P8].

[P1] Mihkla, Meelis; Pajupuu, Hille; Kerge, Krista; Kuusik, Jüri 2004. Prosody modelling for Estonian text-to-speech synthesis. – The First Baltic Conference. Human Language Technologies, The Baltic Perspective, April 21–22 2004. Riga: 127–131.

[P2] Mihkla, Meelis; Kuusik, Jüri 2005. Analysis and modelling of temporal characteristics of speech for Estonian text-to-speech synthesis. Linguistica Uralica, XLI(2): 91–97.

[P3] Mihkla, Meelis 2005. Modelling pauses and boundary lengthenings in synthetic speech. – Proceedings of the Second Baltic Conference on Human Language Technologies, April 4–5, 2005. Tallinn: 305–310.

[P4] Mihkla, Meelis; Kerge, Krista; Pajupuu, Hille 2005. Statistical modelling of intonation and breaks for Estonian text-to-speech synthesizer. – Proceedings of the 16th Conference of Electronic Speech Signal Processing, joined with the 15th Czech-German Workshop "Speech Processing", Robert Vich (Ed.), September 26–28. Prague: 91 – 98, Dresden: TUDpress.

[P5] Mihkla, Meelis 2006. Pausid kõnes. Keel ja Kirjandus, XLIX(4): 286–295.

[P6] Mihkla, Meelis 2006. Comparision of statistical methods used to predict segmental durations. – The Phonetics Symposium 2006: Fonetiikan Päivät 2006, Helsingi, 30.–31.08.2006. (Eds.) Aulanko, Reijo; Wahlberg, Leena; Vainio, Martti. Helsingi: 120–124, University of Helsinki.

[P7] Mihkla, Meelis 2007. Morphological and syntactic factors in predicting segmental durations for Estonian text-to-speech synthesis. – Proceedings of the 16th International Congress of Phonetic Sciences, Saarbrücken, 6–10 August 2007. (Eds.) Jürgen Trouvain, William J. Barry. Saarbrücken: 2209–2212.

[P8] Mihkla, Meelis 2007. Modelling speech temporal structure for Estonian text-to-speech synthesis: feature selection. Trames. Journal of the Humanities and Social Sciences, 11(3): 284–298.

1. INTRODUCTION

1.1. Objectives

An important keyword in speech technology is speech variability. While in speech recognition the variability in the speech wave is a frequent source of problems, an insufficient variability in speech synthesis may lead to monotony and unnaturalness of synthetic speech (Tatham, Morton 2005:9). The naturalness of the temporal structure of the synthetic speech requires a successful rendering of the variability in the duration of speech sounds and pauses, as well as the position of pauses in the speech flow.

The present research was largely motivated by the relative monotony and poor fluency of the output speech of the Estonian text-to-speech synthesiser developed in 1997–2002. The synthesiser was based on a rule-based prosody model (Mihkla, Meister, Eek 2000). Among the shortcomings of the rule-based prosody models is that they rely on the generalisations made on the basis of measuring the so-called laboratory speech, and that these models contain mistakes due to the simultaneous implementation of independently derived rules. The use of large connected speech corpora and statistical optimisation, however, make it possible to replace rule-writing with statistical modelling and to improve the quality of synthetic speech (Sagisaka 2003).

The present dissertation aims to develop a methodology for an automatic modelling of the temporal structure of speech for high-quality text-to-speech synthesis. To this aim, connected speech corpora were subjected to various statistical methods (linear and logistic regression, classification and regression trees (CART), and neural networks) in order to predict durations of speech units (i.e. speech sounds and pauses). These statistical techniques will also be used for generating speech prosody for Estonian corpus-based synthesisers that rely on variable-length speech unit selection algorithms (Mihkla et al. 2007). Corpus-based modelling of the temporal structure of speech is of interest also in phonetics as it enables us to analyse small and hidden yet significant differences in segmental durations which are dependent on parts of speech [P7]. The corpus-based statistical approach is predicted to become most widespread in phonetic sciences as it enables to test various theoretical approaches on large amounts of data and conduct precise analyses, which will provide a statistically sound basis for the functioning of cognitive regulation mechanisms in phonetics.

1.2. Structure of the dissertation

The present dissertation consists of an introductory part and copies of eight articles. The introductory part has seven chapters.

Chapter I introduces the problems and structure of the dissertation and provides a short overview of the publications and the author's contribution to the co-authored articles. It also explains the terms and concepts related to the temporal structure of speech.

Chapter II gives an overview of the strategies of speech synthesis, speech timing theories and selection principles of the factors and features in the modelling of speech timing.

Chapter III contains a brief overview of the previous research on the temporal structure of Estonian speech: the treatment of the quantity degrees, micro-prosodic features of the segments (intrinsic durations), pauses and prepausal lengthening.

Chapter IV describes the data used in the present research.

Chapter V is devoted to the statistical methods used for predicting durations. It also gives an overview of the statistical programme packets used in this research.

Chapter VI provides a description of the results of a number of modelling experiments, including the prediction of the duration and location of pauses in speech flow. Significant features are selected for the modelling of segmental durations and related issues of word prosody are analysed. Various statistical models are described and the relevance and predictive precision of models is tested. Finally, a comparison of methods for modelling segmental durations is presented.

Chapter VII contains conclusions and directions for further research.

1.3. Brief overview of the articles and the author's contribution to the co-authored works

The present dissertation consists of eight scientific articles. The following is a brief overview of the articles and of the author's contribution to the co-authored works. The co-authors of [P1], [P2] and [P4] have seen and accepted this overview.

[P1] deals with issues related to the modelling of prosody for the Estonian text-to-speech synthesiser: modelling the intonation of questions with *kas*-particle, initial notes on how pauses and prepausal lengthening are related to the text structure and the first modelling of phone durations by using regression analysis. The author wrote the analysis of pauses and prepausal lengthening, prepared the modelling data and interpreted the results.

[P2] presents the statistical modelling of segmental durations for speech synthesis, using regression analysis. The author wrote the analysis of pauses and established the link between pauses and the text structure. The author also contributed to the preparation of the material for the regression model and gathering expert opinions on significant features as well as presenting them in the context of regression analysis.

[P3] concentrates on the analysis of pauses and prepausal lengthening in connected speech and the modelling of pauses and their location in the speech flow. The author wrote the article and carried out the experiments. Jüri Kuusik consulted on the application of logistic regression on the input data.

[P4] is on the modelling of intonation based on morphological, syntactic and parts-of-speech features by using the linear regression method, and provides an analysis of pauses and breathing in speech. Pauses are treated as units which mark the boundaries of prosodic groups. The author focused on the theory, the statistical modelling of the fundamental frequency and the related analysis of the speech material. Hille Pajupuu analysed the pauses and breathing in the speech flow and determined sentence stresses. Krista Kerge carried out a syntactic analysis of the sentences and interpreted the generated models.

[P5] contains a longer treatment of pauses in Estonian speech as well as the modelling of pause durations based on the classic regression analysis, the classification and regression tree (CART) method and neural networks. Logistic regression was used to predict the location of pauses.

[P6] provides a comparison of various statistical prediction methods (linear regression, CART method and neural networks) in terms of their predictive error, model interpretability, preliminary data processing and other criteria.

[P7] investigates whether predicting durations in Estonian, which has a rich morphology, is in addition to morphological information also facilitated by the information on parts of speech and syntax.

[P8] focuses on the selection principles of significant features for modelling the temporal structure of speech for text-to-speech synthesis. In addition to traditional parameters describing the phonetic environment of sounds and its hierarchical position in a clause, morphological, syntactic and lexical features of words such as word form, part of sentence and part of speech also play an important role in predicting segmental durations in the Estonian language. Significant features in predicting the positioning of pauses in the speech flow were the distance of words from the beginning of the sentence and from the previous pause, the duration and quantity degree of the previous foot and the punctuation marks or conjunctions in the text.

1.4. Terms and concepts used in the dissertation

The main aim of the functioning of a **language** as a sign system is to ensure the expression of thoughts and the communication and reception of information by means of spoken language or written text. Language is a sign system used in speaking (**speech**), writing (written language), and thinking (inner speech) or in other form of communication. The ability to speak is not innate; it is acquired through human activity. The biological linguistic abilities of humans have provided a basis for acquiring a language system from speech and using the acquired system when speaking (Õim 1976).

Thus it can be said that linguistic communication is the transmission and reception of thoughts via speech signals. Unfortunately, computers are not yet able to think independently. **Speech synthesis** or, to be more precise, **text-to-speech (TTS) synthesis** is the ability of a device or a computer to translate orthographic text into speech without human interference.

Phonetics studies the expression of linguistic signs in the form of spoken language. The main unit of phonetics is the phone (or speech sound) which is the smallest speech segment that can be determined by articulatory and acoustic properties. Yet a phone has a large number of variants in an acoustic space depending on its context in the word and the speaker. By systematically grouping the phones we can establish the **phonological system** of a language the units of which are **phonemes** (Hint 1998). Thus, the input of speech synthesis is a sequence of text or phonemes which in the output is realised as a sequence of sounds, i.e. **synthetic speech**. In speech recognition, the process is reversed – by analysing speech waves we try to establish the in-depth structure of sounds, i.e. the sequence of phonemes. Kalevi Wiik has compared the relationship between a phoneme and a phone with the situation of a shooter in a shooting range (Wiik 1991): just as a shooter is trying to aim at the centre of the target, a speaker is trying to achieve the same target value of the phoneme /a/ in words such as, for instance, *sada, tanu, pali*, but due to the coarticulatory environment the result is, similarly with the shooting target, not a sound of the exact same quality but a cluster of similar sounds. Smaller linguistic units – **segmental phonemes** – are described through both qualitative properties of sounds as well as a parameter related to the temporal dimension – **intrinsic duration**.

In the presentation of speech (but also in music), a certain order is vital which appears in longer passages of speech than sounds (phonemes). This order is rendered through changes in the duration, fundamental frequency and intensity of the physical parameters of the sound signal. This is what **prosody** deals with. **Suprasegmental phonemes** or **prosodemes** which normally accompany several segmental phonemes can be described by the prosodic features which are formed on the basis of the duration, pitch and loudness (or their various combinations) of the psycho-acoustic perception parameters

derived from physical values. The ability of prosodemes to differentiate meanings is above all based on the distinctive difference of prosodic properties characterising the whole unit. Depending on the nature of the suprasegmental or prosodic phenomenon, the speech segment constituting a prosodeme may be a syllable, foot, word, word combination or sentence. Prosodic phenomena include, for instance, word stress, phrase stress, contrastive stress (focus), syllable tones (e.g. in Chinese), tonal word accents (e.g. in Swedish), Estonian quantity degrees, sentence intonation, etc.

The physical parameter **duration** marks the time spent on pronouncing any speech unit (sound, syllable, foot, word, phrase, sentence, pause, etc) or its part. Duration may depend on the qualitative properties of a given unit (e.g. intrinsic duration dependent on phone quality) or its neighbours as well as on their quantity, position in the word and sentence and many other morphological, syntactic and paralinguistic factors (Eek, Meister 2003). The length of a speech unit is usually perceived as its **duration** (e.g. as a short or long sound).

Fundamental frequency (F0) and its variability (i.e. different fundamental frequency contours) is created by the vibration of vocal cords while articulating voiced sounds, which the listener perceives as pitch or a change in pitch. The F0 flow in a phrase or sentence forms the intonation of this phrase or sentence. The pitch and/or its variability in a syllable characterises syllable tones in a foot, i.e. the tonal word accents. **Intensity** is an energetic speech wave parameter expressing atmospheric pressure differences occurring as a result of the interaction of the lungs and vocal cords, as well as the intensity level of the articulation which the listener perceives as the loudness of the signal.

Stress is a complex hierarchical prosodic phenomenon which, depending on the phonological system of the language, is characterised by various physical parameters (duration, F0, intensity, and also vowel quality). **Word stress** is, depending on the language, either phonological (e.g. in English and Russian) or non-phonological (e.g. in native Estonian words stress usually functions as a boundary marker). Longer words have several stresses, the strongest of which is called the **primary stress** and weaker ones are called **secondary stresses**. In native Estonian words the primary stress usually falls on the first syllable of the word. A foot consists of a strong (stressed) and weak (unstressed) syllable. A foot can also have a third syllable if it ends with a short vowel or, at the end of a word, also with a short consonant. In monosyllabic Estonian words the weak part of the foot is made up of a so-called virtual syllable which is expressed by the word-final lengthening. On a higher level, i.e. in words stressed in a phrase or sentence, different types of contrastive stress usually fall on the foot carrying the primary stress in this word (Eek, Meister 2004). In Estonian, word stress is expressed by the higher F0 of the stressed syllable of the foot as compared to that of the unstressed syllable (Eek 1987). Contrastive stress is distinguished from word stress by a significantly higher F0 of the stressed syllable of the foot

carrying the primary stress (Asu 2004). Alternation of stresses creates the speech rhythm.

Estonian quantity degrees are a prosodic phenomenon. **Quantity degrees** are independent distinctive prosodic units manifested over a disyllabic metric foot consisting of a stressed and unstressed syllable. Their distinctness depends on the duration ratios of adjacent phonemes and differences in F0 contours (and maybe also intensity, vowel to consonant transition, and vowel quality) (Eek, Meister 2004).

2. AN OVERVIEW OF SYNTHESIS STRATEGIES AND MODELS OF THE TEMPORAL STRUCTURE OF SPEECH IN TEXT-TO-SPEECH SYNTHESIS

Text-to-speech synthesisers build on the analogue of a human reading. Figure 1 presents a simplified scheme of reading out loud and the physiological speech organs involved in the reading process.

A human being acquires reading skills during their first decade of life. Thereafter their reading skills continue developing and improving. Having acquired the reading skills, they become automatic. Looking at reading from the physiological point of view, we can see that it is a very complicated process. Images of letters are grasped by the sensor neurons of the eyes and transported to the human brain in the form of electrical stimuli. In the brain, the information is processed and translated into commands to motor neurons responsible for activating lungs, vocal cords and articulation muscles (Holmes 1988). This leads to the production of speech, whereas the articulation process is constantly monitored and controlled with the help of the information coming mostly from the auditory organs.

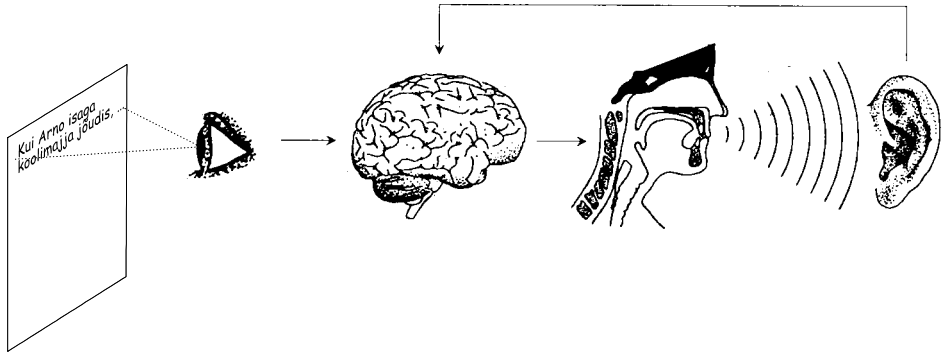


Figure 1. Schematic data flow diagram illustrating the reading process according to Holmes (Holmes 1988).

2.1. Synthesis strategies

The computer imitated TTS system is a simplified model of the physiological reading process (Figure 2).

Similar to a human reading, a TTS synthesiser contains a processing module for natural speech which transforms the input text into the output text together with the desired intonation and speech rhythm. The digital signal processing

module turns the symbol information contained in the input text into natural-sounding speech.

The processing module for natural speech provides the text with a phonetic description and determines the speech prosody. Normally, text processing includes various description levels: phonetics, phonology, morphology, syntax and semantics.

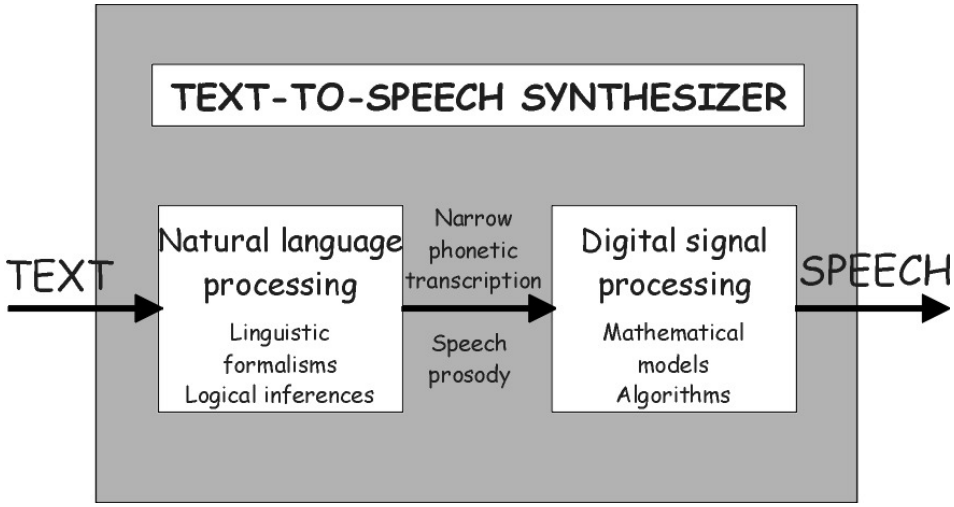


Figure 2. Generalised text-to-speech synthesis model.

In the 1960s speech synthesis techniques became divided into two paradigms. Lingaard called these the system and the signal method (Lingaard 1985). The system method is also called articulatory synthesis. Articulatory synthesis is based on the physiological model of speech production and the physical description of sound production in the speech tract. Both methods developed independently, but the fastest practicable results were achieved by signal modelling thanks to the intrinsic simplicity of the approach. Contrary to the articulatory approach, signal modelling does not even attempt to explain the impact of coarticulation based on the kinematics of speech organs but simply describes the respective acoustic waveforms.

To produce comprehensible and natural output speech, the focus is on modelling transitions from sound to sound and coarticulation. Speech scientists have established long ago that phonetic transitions are just as important for comprehensibility as stationary parts (Liberman 1959). Taking into account phonetic transitions in synthesis can be achieved in two ways: directly – as a list of rules formally describing how phonemes affect each other – or indirectly – by saving phonetic transitions and thereby coarticulatory impacts into a database of speech segments and using them in the synthesis as final acoustic units instead of phonemes.

The two above-mentioned alternatives have developed into two main TTS system types – rule-based synthesis and chain synthesis. Both have a synthesis philosophy of their own.

Rule-based synthesisers are favoured by phoneticians and phonologists because they can be used to study pronunciation mechanisms. The most widespread is the so called Klatt's synthesiser (Klatt 1980) because, due to a link between articulatory parameters and inputs of the Klatt model, it is possible to use the synthesiser in speech physiology research. Unlike rule-based synthesisers, synthesisers based on linking speech units have very little information about the data they operate with. Most of the information is contained in segments which are linked in the chain.

In chain synthesis it is presumed that articulated speech flow is not simply a sequence of phones but rather that speech consists of constantly overlapping transitions from one phone to another. The preceding segment contains, due to regressive coarticulation, features of the following phone. Diphones[5] are the most-used speech units in chain synthesis, as a relatively small number of diphones is needed to synthesise speech from a random text. The Estonian diphone database contains approximately 1,900 diphones. While in common TTS diphone synthesis the speech database contains only one sound-to-sound transition, in corpus based synthesis the whole corpus constitutes the acoustic basis of synthesis. Diphones are also used as elementary units in the corpus-based synthesis of variable-length speech units (Clark et al 2007). Speech unit selection algorithms start their search on the higher levels of the phonological tree (phrase, word, foot), giving preference to longer passages in synthesis.

The primary focus of the present dissertation in modelling the temporal structure of speech is both on TTS chain synthesis based on single diphones (Mihkla, Meister 2002) and on corpus-based synthesis system of unit selection (Mihkla et al 2007). Because diphones contain transitions of adjacent phones, it is wise to treat segmental durations of sounds and pauses as elements of the temporal structure of speech.

2.2. Speech timing

Broadly speaking, there are three different types of timing in speech: mora-timed rhythm which is used to explain, for instance, Japanese, syllable-timed rhythm which is most characteristic of French and Spanish, and stress-timed rhythm which is used in the temporal regulation of many Indo-European languages.

[5] Diphones begin in the centre of the stable part of a speech sound and end in the stable part of the next sound.

In Japanese, mora isochrony has been observed as a temporal constraint controlling vowel duration. A negative correlation has been found to exist between the durations of vowels and their adjacent consonants. The phenomenon by which the temporal compensation of the duration of a vowel is more influenced by the duration of its preceding consonant is regarded as an acoustic realisation of mora-timing. Statistical analysis has shown that such compensation takes place in mora units and not in syllables (Sagisaka 2003). Mora metrics has been successfully applied in Estonian phonology as well. Arvo Eek interpreted intra-foot quantity degrees as a manifestation of mora isochrony where the quantity degree is determined by the distribution of durations within the foot (Eek, Meister 2004:336–357).

In a syllable-timed language, every syllable is thought to be roughly of the same duration when pronounced, although the actual duration of a syllable depends on the situation and context. Spanish and French are commonly quoted as examples of syllable-timed languages though there is no consensus in this respect (e.g. Wenk, Wioland 1982). When a speaker repeats the same sentence several times at the same rate of articulation, the durations of adjacent phones display a strong negative correlation, i.e. any variance in the duration of a single phone is compensated by the duration of adjacent phones. Thus the temporal regulation of articulation must be organised at levels higher than the phoneme, e.g. at the level of the syllable (Huggins 1968). The hypothesis of syllable-timing was applied by Campbell and Isard in the statistical modelling of the interaction between higher and lower levels (Campbell, Isard 1991).

In a stress-timed language, syllables may have different durations but the mean duration of the stretch between two consecutive stressed syllables is more or less constant. Isochrony has been under careful scrutiny in many languages for a long time; yet there is no consensus about speech timing and its acoustic features. In her extensive study (Lehiste 1977) of isochrony and speech rhythm, Ilse Lehiste concluded that the English language lacks direct acoustic correlates related to speech rhythm. It seems that Thierry Dutoit was probably right in saying that there are no so called “clean” languages that would completely match one of the above mentioned rhythm models, and it is more appropriate to talk about tendency to isochrony in languages (Dutoit 1997). In recent studies of the Estonian quantity system, it is considered appropriate to describe the quantity degrees in the context of foot isochrony (Wiik 1991; Eek, Meister 2003).

At the International Congress of Phonetic Sciences held in Saarbrücken in 2007 an entire session was devoted to speech timing with scholars of various languages (English, Japanese, Brazilian Portuguese and French) discussing the mechanisms of rhythm. Although there was not a complete consensus among scholars, many of them focused on different aspects of vowel onsets in the temporal structure of speech (Keller, Port 2007). The onset of voicing has provided a key for studying the temporal structure of syllables. Vowel onsets

play an important role in making speech synthesis more natural and contain significant parameters for speech perception (Keller 2007). Curiously, the new approach discussed at the conference greatly resembles the foot theory of Estonian quantity degrees where the relationship between the duration of the rhyme of the stressed syllable and the duration of the core of the unstressed syllable has an important role to play[6].

Estonian has characteristics of a stress-timed language. In the present dissertation, the modelling of speech duration takes into account the treatment of Estonian quantity degrees and stress which account for the main features of the Estonian syllable and foot structure.

2.3. Statistical methods in prosody modelling

Science follows technology and sometimes technological constraints narrow scientific approaches (Campbell 2000). Twenty years ago, when oscillograms and spectrograms were used to measure durations the size of the paper used for printing placed restrictions on the duration of the sample under study. For this reason, earlier research was mostly based on the study of words or phrases presented in short frame sentences. Since the volume of analysis was restricted, the focus was on "laboratory speech" in which segment durations may considerably differ from those measured in connected speech (Campbell 2000). Later, when automatic analysis and processing of large speech databases became possible, the temporal structure of speech was studied on the basis of connected speech. Another reason why statistical modelling was adopted lied in rule-based prosody systems.

Rule-based timing models were able to determine segmental durations in most cases, but sometimes serious errors occurred, often due to an attempt to simultaneously apply independently derived rules. When large speech databases became available, however, they helped to avoid the occurrence of errors in rule-based modelling and to determine durations more precisely while applying statistical procedures.

Duration prediction is a challenge for both mathematicians and linguists. The pioneer of statistical duration modelling was Michael Riley who in 1989 described the application of the CART method (classification and regression trees) for the prediction of segmental durations (Riley 1989). CART uses data to generate a binary tree, recursively classifying data and minimising error variability. Since then numerous studies have been published on the use of statistical methods to predict speech unit durations in many languages. Nick Campbell was the first to use neural networks for calculating syllable durations

[6] The quantity degree in the foot is defined as $\sigma_{stressed}$(nucleus+[coda]) / $\sigma_{unstressed}$(nucleus).

based on context. The Japanese have, in principle, remained faithful to the regression models (Kaiki et al 1992; Sagisaka 2003). Irrespective of the prediction technique applied, statistical modelling has several advantages as compared to rule-based systems.

The first one is its precision and clarity. Statistical optimisation averts major errors which are caused by, for example, unpredictably poor combinations of timing control rules. Moreover, statistical techniques enable to analyse small hidden yet significant differences [P7]. Diminishing the occurrence of major errors is certain to improve the naturalness of synthetic speech and the options of precise analysis provide a good overview of the regulation models in phonetics (Sagisaka 2003).

Another advantage of corpus-based modelling is its scientific bas. In many cases rule-based synthesis lacks clear data description, control algorithms and error measurement options. Corpus-based statistical modelling enables us to find out the exact boundaries of timing control and receive information for improving it by changing the corpus, control algorithms or error measurement. This gives us a systematic scientific method for developing empiric rule-based applications on the basis of error analysis. Such a corpus-based statistical approach is hoped to become most widespread in phonetic sciences where every theory is normally tested under various circumstances, using different data sets and measurements (Sagisaka 2003).

In the present dissertation, the author resorts to various statistical methods (linear and logistic regression, neural networks and CART) in modelling the temporal structure of speech based on text and speech corpora.

3. STUDIES AND MODELLING OF THE TEMPORAL STRUCTURE OF ESTONIAN SPEECH

A number of studies of the temporal structure of Estonian speech have been published which either attempt to describe the durational structure of speech or offer a comprehensive overview of the phenomena of speech prosody based on experimental phonetic results.

In her doctoral dissertation, Taive Särg presents a detailed description of the development of the prosody of the Estonian language (2005): "On the basis of the studies from the 17th to 19th centuries which deal with linguistics and poetry it can be said that prosodic features distinguishing the meaning of words and those significant in terms of the form of the language and folk song only started to be recognised in Estonian." Back then prosody descriptions were influenced by theories that had been developed studying Indo-European languages, and a major contradiction from the point of view of Estonian lay in that until the end of the 19th century scholars failed to distinguish between stress and duration (Preminger, Brogan 1993).

When the writings on phonetics published in the first half of the 20th century focused on describing the correct Estonian pronunciation, phonetic research conducted in the second half of the century already made use of the technology of experimental phonetics and later also computers. Contemporary research on the temporal structure of Estonian speech based on objective measurements starts in the 1960s (Lehiste 1960; Liiv 1961; et al). The following is an overview of the studies of the durational structure of Estonian speech carried out in the framework of experimental phonetics.

The treatment of the temporal structure has focused more on the quantity system (i.e. the quantity degrees) than segmental durations. In the Estonian prosody, contrastive use of duration is recognised. Contrastive quantity degrees in Estonian are short, long and overlong, marked as Q1, Q2 and Q3 respectively. The Estonian quantity degrees enable to express lexical and grammatical differences through quantity only, without having to change the segmental structure of the word (e.g. *jama, jaama* Gen, *jaama* Part; *suga, suka* Gen, *sukka* Part).

In Estonian there are 9 vowel phonemes and 17 consonant phonemes. All vowels can occur in the three contrastive quantities in the first syllable of a word, just as almost all consonants can occur in the three contrastive quantities on the border of the first and second syllable. According to the measurements carried out by Ilse Lehiste, the average vowel durations in first open syllables in the three quantities are 110, 180 and 230 ms, the approximate ratio being 2:3:4 (Lehiste 1960). In the perception of the linguistic quality, it is the duration ratios of the segments in the foot that play a more important role than segmental durations.

Table 1. Duration ratios of the stressed and unstressed syllables as measured by different researchers.

	Q1	Q2	Q3
Lehiste 1960	0.7	1.5	2.0
Liiv 1961	0.7	1.6	2.6
Eek 1974	0.7	2.0	3.9
Krull 1991,1992	0.5–0.7	1.2–2.1	2.2–2.9
Alumäe 2007[7]	0.6–1.0	1.5–2.6	2.1–4.0

The Estonian prosodic system is hierarchical: segment (phoneme), syllable, foot, word, phrase, and sentence. The crucial question is on which level of hierarchy is it most practical to describe the quantity phenomena? As the once proposed segmental quantity theory has gained little support, most scholars have considered the domain of the quantity degrees to be the syllable (Hint 1997; Viitso 2003) or the foot consisting of a stressed and unstressed syllable (Wiik 1985; Eek, Meister 1997; Lehiste 1997; Ross, Lehiste 2001). Duration measurements have shown that the quantities are characterised by a certain duration ratio of the stressed and unstressed syllables in the foot (Lehiste 1960, Eek, Meister 1997). Table 1 presents the duration ratios of the stressed and unstressed syllables measured in a foot by different researchers.

Earlier studies on quantity are mainly based on laboratory speech (isolated words, words presented in frame sentences or isolated sentences). However, Diana Krull showed that such characteristic ratios are also preserved in spontaneous speech (Krull 1997).

On the basis of the study of a tempo-corpus, Arvo Eek and Einar Meister propose new phonetic correlates for the classification of quantity degrees to replace syllable duration ratios. Instead of contrasting syllable and foot quantity theories, they suggest: “It is pointless to make a distinction between syllable and foot quantities, especially while the three-way opposition of syllable quantities is viewed within a foot and a quantity is not recognised from a stressed syllable but with the help of the phonetic properties of the foot. Therefore it makes more sense to simply talk about quantities.” (Eek Meister 2003)

Although duration ratios play an essential role in the perception of the quantity degrees, the fundamental frequency is also important in distinguishing between, e.g., Q2 and Q3 (Lehiste 1960; Liiv, Remmel 1975; Eek 1987). In speech prosody, the durational structure of speech often needs to be viewed together with the fundamental frequency and intensity. A thorough treatment of Estonian sentence intonation is presented by Eva Liina Asu in her doctoral dissertation (Asu 2004).

[7] These ratios are calculated on the basis of automatically labelled connected speech (see http://keeletehnoloogia.cs.ut.ee/konverents/slaidid/alumae.pdf)

In predicting segmental durations it is important to know the intrinsic durations of the speech sounds and the influence of adjacent phonemes. Intrinsic durations and the coarticlation of speech sounds have been studied in many languages. Those universal linguistic phenomena are also present in the Estonian language. Intrinsic durations of Estonian vowels were first measured about half a century ago (Liiv 1961). In several subsequent studies of microprosodic variations in Estonian speech sounds it has been found that in Estonian, short open vowels are about 10–15 ms longer than high vowels (Eek, Meister 2003:836; Meister, Werner 2006:111). Coarticulation includes such phenomena as the shortening of consonants in consonant clusters, in particular in the environment of voiceless consonants (Eek, Meister 2004:267).

Pauses and final lengthening in Estonian speech have been studied only in passing in the context of other tasks. Ilse Lehiste checked whether final lengthening correlates with following pauses and was able to establish a very weak link (Lehiste 1981). Diana Krull viewed prepausal lengthening in disyllabic words in dialogue in relation to quantities (Krull 1997). Arvo Eek and Einar Meister measured sentence-final lengthening on the basis of a tempocorpus (Eek, Meister 2003) but they only viewed words with certain structure, and their main focus was on quantity features. Therefore it was necessary to measure pauses and final lengthening in connected speech for the Estonian TTS synthesis.

One of the first scholars to attempt the modelling of Estonian quantities in the form of a sequence of rules was Kalevi Wiik. He used mora metrics to present Arvo Eek's quantity measurement data and built his synthesis rules upon them (Wiik 1985). In the 1980s, the Estonian Cybernetics Institute developed several prototypes of parametric speech synthesisers. Those synthesisers were also provided with rule-based prosody models controlling the temporal structure and intonation of synthetic speech (Meister 1991; Siil 1991).

In 1997–2002 the prototype of the Estonian TTS synthesiser was created. The synthesiser was based on diphones and a rule-based prosody model (Mihkla et al 2000). The rules concerning the temporal structure of speech wave take into account intrinsic durations, duration ratios of quantity in a foot and the main characteristics of Estonian stress and syllable structure. The temporal structure model contains several tables with durations and a large number of rules controlling segmental durations depending on the context. The values of pauses and pre-boundary lengthening are not modelled; they are added to the speech flow as constant values.

This is the first known endeavour to model the temporal structure of Estonian speech with statistical methods.

4. DATA

The aim of the present research is to analyse and model the durations of segments and pauses in connected speech for Estonian TTS synthesis. The source material consists of various types of read texts. A one-to-one correspondence between text and speech enables a transition from a symbolic presentation of prosody to an acoustic one, as well as to find out whether and to what extent the syntactic structure of the written text is related to the prosodic structure of speech.

The source material is comprised of passages from a CD-version of a detective story (Stout 2003) read by a professional actor, passages from longer news texts read by news announcers of the Estonian Radio and passages from the Estonian phonetic database BABEL (Eek, Meister 1999).

In total, 66 passages read by 27 speakers (14 men and 13 women) were analysed. As a rule, the speakers read different passages; only the BABEL recordings contain passages read by 2 to 3 different speakers. The speech material was manually divided into segments and pauses. As passages from the BABEL corpus had already been segmented, the same phonetic transcription system was applied to the rest of the material (Eek, Meister 1999). The total duration of the speech material was 46 minutes of which the longest material (9.25 minutes) came from a female speaker.

It is a well-known fact that segmental durations follow the nominal distribution on a logarithmic scale. Therefore, the logarithmed duration was used in most modelling experiments [P1], [P2], [P3], [P5], [P6], [P7] and [P8] as a function feature (Figure 3). Input, or in other words, argument features were generated on the basis of read texts. To determine compound boundaries, Q3 and palatalisation, a linguistic processing module developed for TTS synthesis was used (Kaalep, Vaino 2001). Syntactical sentence analysis for [P4] and [P7] was manually carried out by Krista Kerge and Katre Õim. Methods developed in the Institute of the Estonian Language (Viks 2000) were used to analyse morphological and part-of-speech information of the words in [P7].

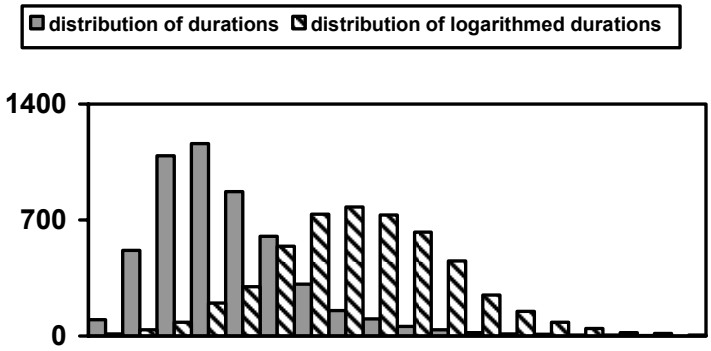


Figure 3. Distribution of segmental durations and logarithmed durations on the basis of the data from one male speaker.

5. METHODS

5.1. Methods and terms of statistical modelling used in the work

The following statistical methods were used in the modelling of the temporal structure of speech: linear regression ([P1], [P2], [P3], [P4], [P5], [P6], [P7] and [P8]); logistic regression ([P3], [P5] and [P8]); classification and regression trees ([P5], [P6] and [P8]); neural networks ([P5], [P6], [P7] and [P8]).

There are excellent descriptions and manuals available on all of these statistical methods, e.g. classification and regression trees (Breiman et al 1984), neural networks (Gurney 1997), linear regression (Weisberg 1985) and logistic regression (Hosmer, Lemeshow 2000). Before moving on to the application and comparison of the methods, let us take a look at the terms used in the dissertation:

Variable – symbol of variable value containing information either in numerical or symbolic format;

Input or argument features – variables used to predict output (in the present thesis it is expected that argument features are determinate and form a vector of argument features $X=(x_1, x_2,\ldots,x_p)$.);

Output or function feature – variable the value of which is calculated on the basis of inputs;

Model – a set of equations or algorithms used for estimating the output value on the basis of input;

Weights – numerical values used in a model;

Parameters – optimal values of weights in a model;

Training – the process of determining the optimal values of weights in a model or the selection of optimal branching variables and values in a tree model;

Training data – input-output data used to estimate weights in training;

Test data – input-output data not used in training;

Validation data – input-output data remotely used in training to select the model or when the training is stopped;

Categorical variable – variable with a limited amount of potential values;

Nominal variable – numerical or symbolic categorical variable in which categories are unordered;

Ordinary variable – numerical or symbolic categorical variable in which categories are ordered;

Interval variable – numerical variable with informative value variations;

Binary variable – variable with only two different values.

5.2. Statistical programmes used

The first measurements of segmental durations were conducted in MS Excel environment, applying the regression analysis tool ([P1], [P2]) of the additional module Analysis ToolPak. The next tool that was used was the statistical programme package SYSTAT 11. This statistical modelling programme enabled us to use multiple linear regression, regression trees and logistic regression for the prediction of pause locations ([P3], [P4], [P5]). In the framework of the doctoral school it was also possible, thanks to the Institute of Applied Statistics of the University of Tartu, to use the statistical programme SAS 9.1. The environment of the programme Enterprise Miner proved to be most convenient for statistical modelling thanks to the possibility of simultaneous application of various methods and comparison of model compatibility and the results of different methods ([P6], [P7], [P8]). The programme was easier to use because in the SAS environment the input data need not be processed in advance (e.g. transforming categorical variables into binary pseudo variables). The processing is automatic. Figure 4 presents a typical data flow scheme in the SAS Enterprise Miner environment used in the research.

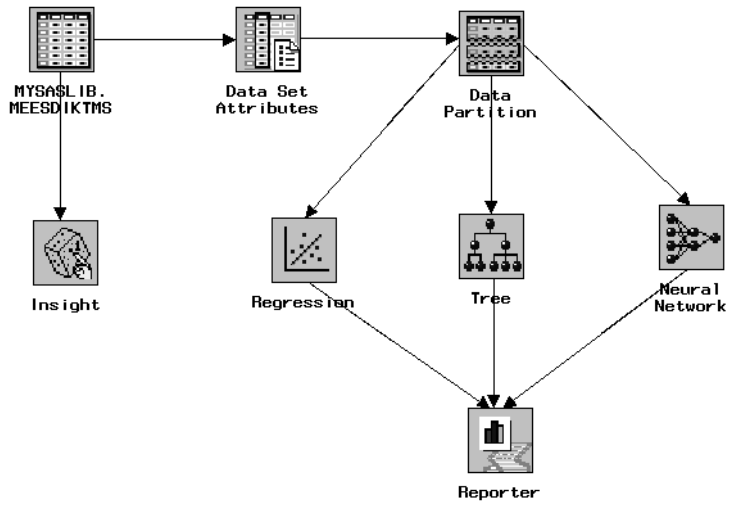


Figure 4. The SAS Enterprise Miner working environment for modelling the temporal structure of speech.

Description of the data flow modules:

MYSASLIB.MEESDIKTMS	– input data of a male radio announcer
Insight	– a good tool for getting an overview of the data by variables; it helps to identify invalid or missing data
Data Set Attributes	– a module for determining variable functions in a model (i.e. which is a dependent variable or function feature and which are model inputs or argument features)
Data Partition	– classification of input data into training, validation and test data
Regression	– regression analysis module
Tree	– decision trees module
Neural Network	– neural networks module
Reporter	– results presentation module

6. RESULTS

6.1. Analysis of the durations and locations of pauses and pre-boundary lengthenings in connected speech

To make synthetic speech sound natural to a human ear, it needs to have natural-sounding intonation, rhythm and stress. In other words, the TTS system needs to be able to generate such durations of segments and pauses and values of the fundamental frequency which would not significantly differ from the respective values in natural speech (Zellner 1994). In phonetics and phonology, relatively little attention has been paid to pauses so far. Linguistic research of spoken language treats speech sounds, syllables, feet, words and phrases as speech units, mostly in the context of isolated sentences. It is, however, difficult to view pauses as functional speech units within a sentence, which might explain their relative neglect in linguistic-phonetic studies (Tseng 2002). In the last decade, however, after speech corpora started to be widely-used in phonetic research, pauses as a significant feature of speech prosody started to receive much more attention.

In the present dissertation pauses are analysed hand in hand with segmental durations ([P1], [P2], [P8]). [P4] views pauses and breathing in connected speech as units marking the boundaries of prosodic groups. [P3] and [P5] are devoted to the analysis of pauses and pre-boundary lengthening and on modelling pause durations and locations in connected speech. While in [P3] only linear and logistic regression is used for modelling, in [P5], which presents a summary of pauses, the CART method and neural networks are also applied to model pause durations.

As the title of Figure 2 in [P5] is incomplete, we would like to specify it here (Figure 5) in order to illustrate the location of pauses in connected speech in Estonian. The left column of the figure contains read text and the right column is a simplified presentation of the respective speech flow – pauses in a sequence of graphemes. We can see that the structure of the text is much stricter: on the whole there is a space at the end of each word and a punctuation mark at the end of each sentence. When speaking, every person is quite free to interpret a text: pauses separating the words follow a word group or a prosodic phrase, but prosodic phrases do not necessarily coincide with syntactic phrases, and prepausal lengthening has a tendency (although not always) to come at the end of a prosodic phrase. Some of the underlined feet in Figure 5 are lengthened due to focus (e.g. in the phrase *veetlevate noorte naiste seltskonnas* ‘in the company of charming young women’ the word *naiste* ‘women’ is highlighted with a foot lengthening).

Talle meeldis nendega uhkustada – kui need teie omad oleksid, meeldiks see teilegi –, aga mitte sellepärast ei seganud ta vahele. Ta tahtis paari kirja dikteerida ja ta arvas, et kui ma missis Hazeni üles orhideesid vaatama viin, siis ei tea keegi, millal me sealt alla tuleme. Aastaid tagasi jõudis ta ebapiisavatele tõenditele tuginedes otsusele, et ma kaotan veetlevate noorte naiste seltskonnas ajataju, ja kui tema kord midagi otsustab, siis on see otsustatud.	Tallemeeldisnendegauhkustada**P**kuineed teieomadoleksidmeeldiksseeteilegi**P**aga mittesellepärasteiseganudtavahele**P**Tatah tis**P**paarikirjadikteeridajataarvasetkuima missisHazeniülesorhideesidvaatamaviins iiseiteakeegi**P**millal**P**mesealtallatuleme**P** Aastaidtagasijõudista**P**ebapiisavateletõe nditeletuginedesotsuseleetmakaotanveetl evatenoortenaisteseltskonnasajataju**P**jak uitemakordmidagiotsustabsiisonseeotsus tatud**P**

Figure 5. The structure of read text versus pauses in the speech flow. The left column presents the read text and the right column shows pauses in the speech flow (**P** – pauses separating words, underlined graphemes – lengthened feet).

In [P3] and [P5], above all such pauses and prepausal lengthening was analysed which were related to punctuation marks and conjunctives. To this aim the duration of pauses was measured on the basis of the speech wave of the read texts and the lengthening of feet was calculated. To calculate the lengthening in feet, durations of the segments forming a foot were added up and the result was compared to the mean duration of a given foot structure in the speech of each speaker. Besides structure, foot quantity was also taken into account. If a foot structure proved unique in the text (e.g. CVCCC-CV word *'korstna'*), its duration was compared with that of a similar foot structure (e.g. CVCC-CV word *'kordse',* subtracting the duration of one component of a consonant cluster from the sum of segmental durations of the word ‚*korstna'*).

Table 1 of [P3] and [P5] presents the mean durations of pauses and prepausal lengthening in the speech of 27 informants. It can be seen from the table that even the mean values have an extreme high variability. It is, however, interesting to note that the general means of pause durations in the material of male and female informants differ only within 10%. Visual observation of the general means suggests that in a text read at a normal speaking rate pauses can be distinguished by their duration. Statistical analysis of the samples also confirms this observation. It is possible to differentiate between phrase-final, sentence-final and passage-final pauses in speech. In the analysis of the foot lengthening data with the Student t-test, we had to maintain the zero hypothesis: cases of foot lengthening came from samples with identical means.

The second stage was to find out to which extent, if at all, the prosodic structure of speech correlates with the syntactic structure of text as marked by punctuation marks and conjunctions. Table 2 of [P3] and [P5] shows that there is a pause at the end of each passage and almost every sentence. Only a

professional actor took the liberty of reading two sentences as one. The analysis of the colon and dash also showed that there is a very strong link between syntax and prosody. Two thirds of commas elicited pauses. The least marked phrases in speech are those beginning with coordinating conjunctions (*ja, ning, ega, ehk, või, kui ka*) which normally do not require a comma.

Of all the punctuation marks it was the dash which had the clearest connection with final lengthening. This is probably also due to the shape of this punctuation mark – a long line makes speakers stretch words. The term "prepausal lengthening" refers to the connection between pauses and final lengthening. In the Estonian speech material, this term applies to only 60% of the cases (of the 601 pauses only 360 were preceded by foot lengthening). Perception tests carried out by Lehiste (Lehiste, Fox 1993) show that as compared to e.g. English speakers, Estonian speakers expect the final lengthening of the last syllable to be considerably shorter.

Our analysis showed that although pauses in speech are very variable, it is possible to distinguished different types of pauses on the basis of duration. This cannot be said about prepausal lengthening. It is doubtful that producing a constant, phrase-final pause after every other comma and every third conjunction would improve the rhythm and naturalness of synthetic speech. Rather, naturalness of synthetic speech can be achieved by our ability to render the variability of pause durations and locations in the speech flow.

6.2. Feature selection for the modelling of segmental durations and expert opinions

In almost all statistical models, the selection of the factors and features of durational models relies, to a greater or lesser extent, on Dennis Klatt's rule-based model (Klatt 1979): speech segments have their intrinsic durations; they are influenced by adjacent segments; segmental duration depends on its location in the syllable, word and phrase but also on the overall context – on the duration of the syllable, word and phrase. In stress-timed languages, syllable stress and the contrastive stress of a word are also important. In addition to general features, durational models also contain specific phonetic information about a language. For example, the prediction model of the temporal structure of German segments contains a syllable structure feature (Möbius, van Santen 1996). Syllable structure is also important in Hindi (Krishna, Taludar, Ramakrishnan 2004). Petr Horák introduced a special feature for monosyllabic words into the Czech durational model (Horák 2005). Similarly with Czech, Dutch has a special feature for clitics and also for word frequency (Klabbers 2000). It is thus presumed that more common words are pronounced slightly differently from those rarely found in texts. In languages with a large number of

function words, a distinction is made between function words and content words (Brinckmann, Trouvain 2003; Klabbers 2000). Martti Vainio included morphological features and part-of-speech information into prosody modelling for Finnish TTS synthesis (Vainio 2001).

Feature selection for the modelling of segmental durations in Estonian was based on the principle that the Estonian stress and quantity degrees are described in the framework of a prosodic hierarchy enabling to divide an utterance into components lying on different levels of subordination (Eek, Meister 2004:253). As can be seen in Figure 6, a sentence or phrase[8] consists of prosodic words, while the words, in turn, consist of feet; the feet consist of syllables and the lowest, segmental level is represented by phonemes. In all studies on the prediction of segmental durations ([P1], [P2], [P6], [P7] and [P8]), the relative position of a speech unit in a sentence is presented in a hierarchic scale as follows: the position of the phoneme in the syllable, position of the syllable in the foot, position of the foot in the word, and position of the word in the phrase. In addition, as has been shown by previous analysis, information describing the levels of prosodic hierarchy is also important: syllable stress, open vs. closed syllable, quantity degree of the foot, phrase length in words, etc. The above-described feature system relies heavily on the parameters of Klatt's rule-based temporal structure model. A special feature of Estonian is the foot as a phonological level. Following the example of the Czech researcher Pavel Horák (Horák 2005), the monosyllabic word feature, which proved to be significant in modelling, has also been added in some of our latest studies ([P6] and [P8]).

[8] In Estonian, phrases (noun, verb and adverbial phrases) are often closely intertwined in sentences, which is why in the present dissertation a phrase is a clause or an element of a list followed by a punctuation mark or conjunction in the sentence. In the example given in Figure 6 the sentence and the phrase are equal.

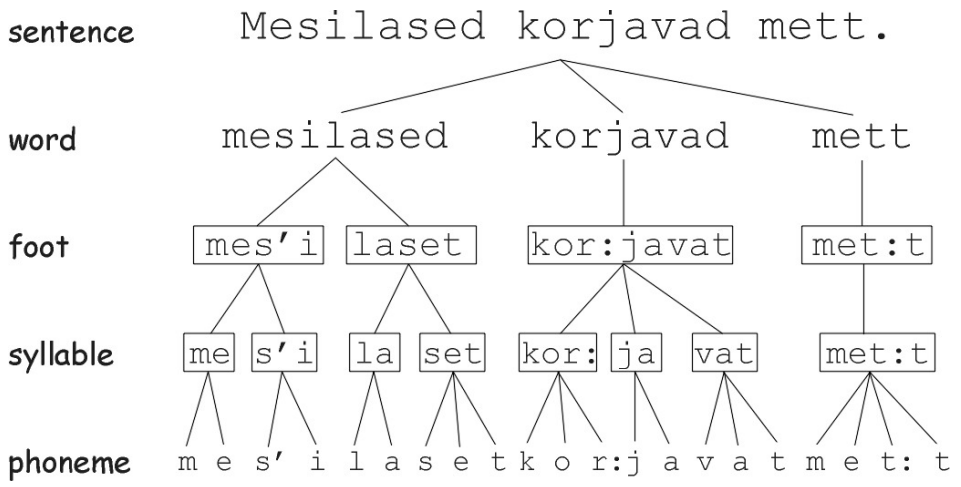


Figure 6. Hierarchical encoding of a speech unit in the phonological structure. For example, the location of the phoneme [l] is encoded according to its position in the two-phoneme syllable [la]. The position of the syllable [la] is encoded in relation to the disyllabic foot [laset] and that of the foot according to its place in the word [mesilased] etc.

The next underlying principle of feature selection is that every phone has its intrinsic duration and that each speech sound is affected by its adjacent sounds. How many adjacent phonemes to the right and to the left affect the duration of a given phoneme? In our first studies ([P1], [P2]) the influence of only one neighbouring phoneme from each side was taken into account. In last experiments ([P6], [P8]) it was considered optimal to view two neighbouring phonemes (i.e. at least the next and next but one on the right and the previous and previous but one on the left, see Figure 7). Phonemes are defined by their class (9 classes, including pauses), and contractive length (short vs. long).

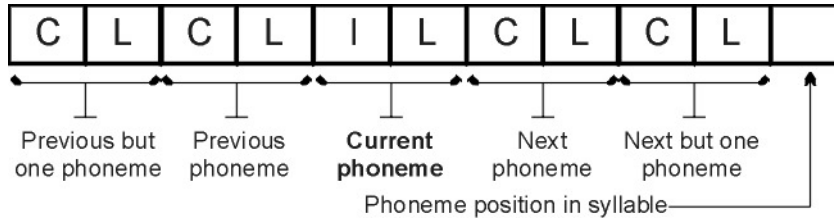


Figure 7. Encoding the information on the location of a current phoneme depending on its context.
(C – phoneme class, L – phoneme contrastive length, I – phoneme identity).

It has proven optimal to describe a phoneme and its context with 10 features, the hierarchical position of the phoneme in the utterance with 5 features, the characteristics of some speech units (syllable stress, syllable type, quantity degree of the foot) with 3 features and the information about the duration of higher-level units (syllable, foot, word, phrase, sentence) with 5 features. In addition, a binary feature is used which refers to punctuation marks after certain words in read text. All these features (24 in total) make up a vector of basic features to serve as input for the durational model [P8]. Another important point to in the selection of initial features was that it had to be possible to generate all the features automatically from the input text. In all the studies dealing with the modelling of segmental duration ([P1], [P2], [P6], [P7] and [P8]) we made use of a sentence builder, syllabifier, morphological analyser, disambiguator and other modules provided by Estonian language technologists (Viks 2000; Kaalep, Vaino 2001).

After an initial selection of the features, it was possible to obtain expert opinions on the vector of the chosen argument features as well as recommendations on which features to add. The experts were asked to estimate whether or not a feature was significant in the prediction of speech timing (e.g. segmental durations) and to give their opinion on the possible joint effects of certain features. During our first experiments in statistical modelling we invited six Estonian phoneticians and speech technologists to evaluate our first vector of argument features. The overlap between their opinions and our preliminary results was a mere 41–65% [P2]. However, as a result of adding more speech material and increasing the volume of Estonian speech corpora our recent results are in better accordance with expert opinions [P8]. The still considerable difference between the two sets of results can be explained by the fact that the so called “duration patterns” of the phoneticians are largely based on measurements of laboratory speech (isolated words and sentences), whereas our results draw on connected speech. Segmental durations measured on isolated sentences differ greatly from the temporal structure of connected speech (Campbell 2000:312–315).

To sum up, up to 24 features are generated for each phoneme from the input text. These features mostly describe a given phoneme and its context, its location in the hierarchical system and properties of higher-level units. When selecting features and establishing connections between them it is advisable to ask experts for their opinion.

6.3. Comparison of the statistical methods used for the prediction of durations

What is a good method for predicting speech prosody? Are there any objective criteria for selecting the best statistical method? These are questions encountered by any researcher attempting to use statistical methods to model the prosody of connected speech. In our first modelling experiments ([P1], [P2]) we mostly used multiple linear regression. Researchers almost always have doubts about whether the method they use is good enough or if there are perhaps better ones. The author of the present research first started to ponder over those questions during a plenary presentation by Yoshinori Sagisaka at the International Congress of Phonetic Sciences held in Barcelona in 2003. Sagisaka described his more than twenty-year long experience in modelling speech prosody where preference had been given to regression analysis (Sagisaka 2003). When looking at various studies carried out in this field of research (Brinkmann, Trouvain 2003; Horak 2005; Krishna, Murthy 2004; Vainio 2001), it can be noticed, however, that neural networks and regression trees are much more widely used in speech prosody modelling than regression analysis methods. Usually no argumentation in favour of the chosen method is given and prediction results are compared with the existing rule-based prosody generator. The choice of the method seems to be pragmatic and dependent on the educational background of the researcher, his or her supervisors and colleagues, the availability of software, and other factors.

Obtaining a licence of the statistical programme package SAS 9.1 provided an excellent opportunity for us to compare different prediction techniques (regression, CART, neural networks) in the prediction of the segmental durations of segments on the same data. Methods were evaluated in terms of the prediction error, model interpretability, preliminary data processing, and other criteria.

The initial data contained the speech material of one male and one female radio announcer. 26 argument features were generated on the basis of the text. To optimise the number of features, a preliminary data analysis was carried out. By linear regression analysis, features significant for the prediction models generated both from the male and female material were selected. All in all, there were 18 such features (see [P6] Table 1).

The function feature of models for all three methods was the logarithmed durations of speech sounds. Although neural networks and the CART method do not directly require a normal distribution of function features, it does enhance the stability of neural networks.

Statistical modelling results are given in Table 2. All compared methods using the same data and argument features yielded a very similar error percentage. Most surprisingly, linear regression had almost the lowest error

percentage. In essence, linear regression should be able to identify only the most direct and obvious relations between the input and output. Although a certain amount of non-linearity is saved in a regression model when logarithming function features and non-linear encoding of input features, more covert connections between input and output are still trusted to be revealed by more complicated non-linear methods (including classification and regression trees and neural networks). Therefore, it can be concluded that the linear regression method, which for some time has been successfully used to process speech waves (Markel, Gray 1976) and is still used in speech analysis and synthesis, is also a reliable method for modelling the temporal structure of speech.

Table 2. Prediction errors and other evaluation criteria of the models.

Criteria		*Neural networks*	*Regression*	*CART*
Prediction errors: – male announcer (mean error 21%)	Training	0.230	0.230	0.250
	Validation	0.243	0.248	0.264
	Testing	0.230	0.232	0.255
– female announcer (mean error 19%)	Training	0.224	0.221	0.230
	Validation	0.221	0.218	0.231
	Testing	0.221	0.217	0.230
Model interpretation		complicated	easy	very easy
Output normalisation		recommended	necessary	unnecessary
Pre-processing of inputs		necessary	necessary	unnecessary
Interactive training		yes	no	yes
Model with missing input values		no	no	yes

The model can be most clearly interpreted on the binary tree, and the impact of the input on duration is quite easy to understand in the analysis of regression coefficients. It is much more difficult to interpret the results of the learning process on neural networks. Linear regression requires normal distribution of the function feature while other methods do not, although normalisation does enhance the stability of the neural networks model. Before statistical modelling based on regression analysis and neural networks, argument features need to be processed. For regression analysis, nominal features need to be replaced by a large number of binary pseudo-features. The input range of neural networks needs to be [0, 1]. The two features at the bottom of Table 2 are more indirect criteria for the evaluation of methods.

Thus it can be said that, in terms of predictive precision, linear regression is comparable with more complicated non-linear methods (CART, neural networks). However, it is regression trees that yield the best interpretation results.

6.4 Lexical prosody

Traditionally the list of factors significantly affecting speech timing includes neither part-of-speech (POS) information nor morphological characteristics (van Santen 1998, Campbell 2000, Sagisaka 2003). This may be due to most studies on TTS synthesis focusing on languages with relatively little morphology. Finnish is one of the few languages boasting a study of the influence of morphological features on the duration of speech units (Vainio 2001). In Estonian, the word has a very important role both in grammar and phonetics, while the morphology is extremely rich. Hence our interest to check whether there are any morphological, lexical, or even syntactic features possibly affecting the temporal structure of Estonian speech [P7]. Probably the most natural way to calculate the impact of morphological, lexical and syntactic features was through an extension of our earlier methodology of statistical modelling in order to see how these features affect the functioning of the durational models. The modelling was done using two different methods – linear regression and the non-linear method of neural networks. To allow for the qualitative assessment of the impact of the factors, variability of the output error was measured. The results demonstrated a decrease in the output error by a couple of percent when some morpho-syntactic and POS information had been added to the input of the model.

As the durational models in [P7] were based on the speech material of only two radio announcers, it seemed a little premature to make generalisation when interpreting the models. It should, however, be mentioned that the most distinct regularities were revealed by a visual observation of the POS regression coefficients in the regression model. Table 3 presents the mean lengthening and shortening of speech sounds by part of speech in the durational models of male and female speech. As we can see, there is more variation in the middle part of the table, while the top and bottom parts of the table are rather similar. Table 3 shows that in proper names sounds are pronounced longer by 5–6 ms on average. The average duration of the sounds for these two speakers was 62.5 and 64.1 ms respectively. Consequently their pronunciation of proper nouns was about 10% longer than that of the verbs. Nouns and adpositions were pronounced with a little longer duration. Surprisingly the segments of the adpositions were longer than the average. Adpositions are classified as function words. In most languages function words are shorter than content words. An Estonian adposition invariably goes together with a noun which often is focused in the sentence, and therefore its longer than average duration may extend to a neighbouring adposition. Ordinal numerals, however, were pronounced with 10% shorter duration, and pronouns and adverbials with about 5% shorter duration than the average. The shortening of ordinal numerals can be accounted for by a large number of dates in the text, which are typically expressed by

ordinal numerals. Reading the relatively long dates of the past century, the speakers tend to hurry because usually only the last one or two numbers of the year are important, but nevertheless the whole number has to be pronounced, as required by rules of correct reading.

Table 3. Average lengthening-shortening values (in ms) for different parts of speech.

Part of speech	*Male speaker*	*Female speaker*
Proper noun	6.23	5.22
Noun	2.25	2.10
Adposition	0.82	2.82
Genitive attribute	0.42	1.35
Verb	0.00	0.00
Numeral	–0.10	0.42
Conjunction	–0.14	1.81
Adjective	–0.39	1.14
Adverb	–0.89	–2.90
Pronoun	–4.13	–3.86
Ordinal number	–5.44	–7.48

The results demonstrated a decrease in the predictive error by a couple of percent when morpho-syntactic and POS information was added to the input of the durational model. This development was not surprising, considering the important role of the word in Estonian grammar and phonetics.

6.5. Modelling results, significant features, prediction errors and the interpretation of the results

6.5.1. Modelling pauses

To model pause durations, a number of features [P3], [P5] were generated from the text. These features described the text structure (end of passage, end of sentence and end of phrase, conjunctions in the text), prepausal foot (duration of the foot measured in segments, quantity degree of the foot, duration of the last syllable of the foot measured in segments and a binary feature indicating final lengthening); temporal relations of the pause (distance of the pause from the beginning of a passage, sentence and phrase as well as from the previous pause and preceding inhalation).

The predicted feature was pause duration. For the purposes of linear regression, the function feature had to be logarithmed as logarithmed duration can be better subjected to normal distribution.

In the modelling of pause durations on multiple regression the features of the text structure which proved to be significant were the end of passage, sentence and phrase. Of the prepausal foot features, only the binary feature proved significant at the confidence level of 0.05 showing whether the prepausal foot is lengthened or not. In addition, the distance of a given pause from the preceding pause proved to be significant in duration modelling. Figure 8 presents a logarithmed formula for calculating pause duration.

LN(pause duration) *= –1,973 + 0,373 * LQLQP + 1,454 * LALQP + 0,441 * FRKOM + 0,012*KAUGFR+0,024 * KAUGPA + 0,133 *PIKENDUS*

Figure 8. Regression formula for calculating logarithmed pause duration. Variables: LQLQP – end of paragraph feature, LALQP – end of sentence feature, FRKOM – end of phrase (comma), KAUGFR – duration of previous phrase, KAUGPA – distance from previous pause, PIKENDUS – lengthening of the last foot.

In the prediction of pause locations logistic regression was applied in order to predict the probability of a pause following a given word in the speech flow. Largely the same features as for predicting pause durations were used as variables of logistic regression, with only a few exceptions.

The analysis of the speech material of [P4] showed that in addition to punctuation marks and conjunctions, pause locations also tend to correlate with proper nouns and ordinal numbers. However, a subsequent statistical analysis on the basis of a larger set of data showed that such correlations were insignificant. In [P3] the input included two binary features indicating whether the following word is a proper noun or a foreign word. Their addition was inspired by the idea that there might be a short pause before the pronunciation of proper nouns (e.g. *Minu nimi on Tamm, Jüri Tamm.* 'My name is Tamm, Jüri Tamm') and maybe also before some more complicated foreign words (e.g. *Rahvas toetas konstitutsioonilist monarhiat.* 'The people supported constitutional monarchy'). This hypothesis was, however, disproved. The correlation of pauses with proper nouns and foreign words was extremely weak and thus these features proved to be insignificant [P3].

Table 4. Results of logistic regression: variables significantly influencing the location of sentence-internal pauses, the ratio of their odds and confidence levels.

Independent variables	*Odds ratio*	*Confidence levels*	
		Lower	*Upper*
The word is followed by a comma	17.4	11.7	25.9
The next word is a conjunction	7.9	4.8	12.8
Distance of the word from the beginning of sentence	1.1	1.0	1.2
Duration of the preceding foot	1.3	1.1	1.4
Quantity degree of the preceding foot	1.2	1.1	1.5
Lengthening of the preceding foot	6.9	5.2	9.2

Table 4 contains six features which were found by logistic regression to affect the positioning of a sentence-internal pause. A comma in the text is very important, raising the chances for a pause to occur by 17.4 times on average. A word is 7–8 times more likely to be followed by a pause if the following word is a conjunction or the foot of this word is lengthened. Slightly more frequently than average a pause can be expected to occur after longer feet or after words of longer quantity degrees. In a predictive model, however, the role of these features remains relatively marginal because they raise the chances for a pause to occur by no more than 1.2–1.3 times.

In summary it can be said that different types of pauses are distinguishable in speech by their duration. Pause duration and location in the speech flow can be modelled by statistical methods. Due to the great variability of pauses, the generated models describe the mean voice parameters of the informants and only the most general rules in the duration and location of pause. Thus we should gather a sizeable speech material from a couple of speakers and apply the same methods separately to each speaker. It was not possible to create a reliable duration model for the prediction of final lengthening. We will probably need to treat final lengthening as part of the duration model of speech sounds.

6.5.2. Modelling segmental durations

In the durational model, argument features include a number of nominal variables: the identity of a given phoneme (26 phonemes), classes of adjacent phonemes (8 phoneme classes + pause) as well as potential nominal morphological, syntactic and part-of-speech features (see Table 5). Thus there may be up to a hundred regression equation parameters, and it is advisable to interpret the durational model in terms of the significance of argument features. Only in our first studies ([P1] and [P2]) where the number of inputs was smaller and durations were predicted on the level of the class of a given phoneme, we were able to present the results as an equation. The significance of features is most

clearly determined in the regression model where the significance of each feature is given a statistical evaluation. In case of the *forward selection* method, for example, the most significant features of a given moment are being added to the model one by one, with revaluation taking place before each cycle. In the CART-method, too, the most significant features are added to the regression tree. In the method of neural networks, however, there is no such evaluation of significance, and the usefulness of each feature can be estimated by manually adding or removing features while evaluating the output. The results of our first modelling experiments, such as classifying the quantity as an insignificant feature in [P1] and [P2], seemed to be "significant discoveries" in the area of speech prosody. Subsequent experiments ([P6], [P7], [P8]), however, showed that the number of significant features can vary depending on the speaker and that significant features need not always overlap in different methods.

Table 5. Significance of input or argument features for modelling segmental durations.

	Inputs
1.	**Class of the previous but one phoneme**
2.	**Length of the previous but one phoneme**
3.	**Class of the previous phoneme**
4.	Length of the previous phoneme
5.	**Identity of the current phoneme**
6.	**Length of the current phoneme**
7.	**Class of the next phoneme**
8.	Length of the next phoneme
9.	**Class of the next but one phoneme**
10.	**Length of the next but one phoneme**
11.	**Position of the phoneme in the syllable**
12.	Syllable stress
13.	Syllable type
14.	Quantity degree of the foot
15.	**Position of the syllable in the foot**
16.	**Length of the foot in syllables**
17.	Position of the foot in the word
18.	**Length of the word in feet**
19.	**Monosyllabic word**
20.	**Position of the word in the phrase**
21.	**Length of the phrase in words**
22.	Length of the sentence in phrases
23.	**Punctuation**
24.	**Morphology**
25.	**Part of speech**
26.	**Syntax**

Table 5 contains features that on the basis of extensive experimental material have been found to be significant in predicting segmental durations [P8]. The features in bold on a dark grey background were significant for most speakers (over 80%). The features on a lighter grey background were insignificant in the durational models of some speakers (under 50%) Surprisingly the latter group includes the quantity degree of the foot, which is considered a cornerstone of Estonian speech prosody. One of the possible reasons might be that the quality degree as a suprasegmental feature cannot be presented as a single linear feature. Quantity degrees as contrasting quantity models are related to a passage starting with the vowel of the stressed syllable of the foot and ending with the vowel of the unstressed syllable (Eek, Meister 2004). This means that speech contains a large number of sounds (at the onset of the stressed syllable, in the coda of the unstressed syllable, etc) which to not participate in the opposition of quantity degrees. In predicting the duration of such speech units the quantity degree of the foot is probably an insignificant feature. A factor that should probably be included in the hierarchical system describing the positioning of speech sounds is the level characterising the structure of syllable. Syllable structure is reflected in the durational models for German (Möbius, van Santen 1996) and Hindi (Krishna et al 2004). A possible interaction between stress and syllable structure is referred to by the fact that syllable stress does not always have a significant influence on the predicted duration. Surprisingly, the contrastive duration of the next sound is less significant than the duration of the next but one sound [P8]. The two features on a white background in the table, syllable type (open vs. closed) and the position of the foot in the word, consistently proved to be insignificant in the prediction of segmental durations.

6.5.3. Significance of the models and predictive precision

The success of modelling can be evaluated firstly on the basis of the significance of the generated model, and secondly, on the extent of the variability of function features that the model is able to describe. The statistical significance of the whole model needs to be taken checked in the decision process. In sum it can be said that all models generated in the experiments of the dissertation proved to be statistically significant. Table 6 shows an example of the summary of fit and variability analysis of our regression model of pause durations. We can see that the model is statistically significant and the linear relation between the 9 inputs and pause duration is reliable (the probability of model significance due to chance is almost zero). The correlation is relatively strong ($r = 0.83$), which accounts for 2/3 of the variability of pause durations (determination coefficient $r^2 = 0.67$). Other durational models of pauses described 65–73% of the variability of durations, while in the model for segmental durations the number remained in the range of 52–63% ([P6], [P7], [P8]), which is also a

good result. The modelling of final lengthening failed as the generated models accounted only for 25–30% of the variability [P5].

Table 6. Summary of fit and variability analysis of the regression model of pause durations.

Summary of fit					
Mean of response 0.82673			R-square 0.6686		
Analysis of variance					
Source	DF	Sum of squares	Mean square	F stat	Pr > F
Model	9	478.5	18.403	220.87	<0.0001
Error	560	408.8	0.0862		
C Total	559	787.2			

The total predictive precision of the model for the location of pauses reached 88%. Our analysis showed that certain markers of text structure (end of paragraph, end of sentence, colon, dash) are followed by a pause with a 93–100% probability. Leaving out such "well-marked" pauses, the model was able to predict pauses in only 44% of the cases [P5].

Depending on the speaker and the method, the prediction error of segmental durations was within 16.1–21.2%. Testing different methods (linear regression, CART, neural networks) on the same data set showed that the predictive precision of linear regression and the neural networks was nearly equal, while the CART model had a slightly higher predictive error [P6]. It was surprising to find that the linear method could compete with non-linear ones, despite the fact a linear model is usually expected to show nothing but the most obvious and most general relations between input and output, and only non-linear methods are trusted to reveal more covert relations. Figure 9 presents the real segmental durations in a passage of speech as compared to durations predicted on the neural networks (Fišel, Mihkla 2006). In [P5] the error percentage of the general model of pause durations was mistakenly indicated to be 29–37%. The actual error percentage was half as low, i.e. 14–18%. Depending on the speaker, the prediction error for pause durations was 8–12% for the model generated on the basis of speech from one informant.

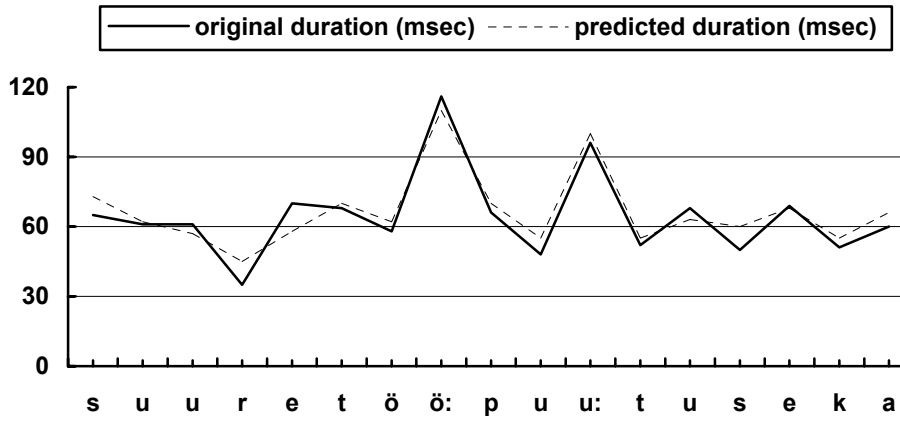


Figure 9. Actual segmental durations vs. durations predicted with the neural networks model.

As a conclusion it can be said that the durational model shows a very strong correlation between output and argument features. There is a strong correlation between pause locations and text structure: punctuation marks and conjunctions. We failed to create a reliable model for the prediction of pre-boundary lengthening. It was also difficult to predict the intra-phrase locations of pauses.

The prediction model of segmental durations contains a relatively large amount of factors and features. It is optimal to describe the context of the current phoneme with two neighbouring phonemes from both sides. Significant features are the position of the phoneme in the hierarchical structure of the sentence and the characteristics of higher phonological levels (syllable, foot, word, phrase), text structure and morpho-syntactic and part-of-speech information.

All models are statistically significant and account for 65–73% of the variability of pauses and 52–63% of the variability of the duration of segmental phonemes.

7. CONCLUSION AND FUTURE RESEARCH DIRECTIONS

The present dissertation is merely the beginning of a long road leading to the creation of a model of the temporal structure of Estonian speech which would include all factors. The main contribution of this research lies in the development of a methodology based on different statistical methods and speech corpora for the modelling and study of the prosody of the Estonian language.

The following results obtained during numerous modelling experiments and statistical analyses could be highlighted:

- during research a corpus of connected speech consisting of texts read by 27 speakers was compiled;
- in the read texts, pauses can be classified as paragraph, sentence and phrase-final depending on their duration;
- durations and locations of pauses in the speech flow can be predicted whereas the strongest correlation was found between pauses and text structure (punctuation marks, conjunctions) and also with the distance from the previous pause and the beginning of the sentence;
- significant features in the prediction of segmental durations were those describing the influence of neighbouring phonemes (two preceding and two following phonemes), the hierarchical position of the sound in the phonological structure of the utterance (the position of a phoneme in the syllable, the position of a syllable in the foot, the position of a word in the phrase, etc) and features characterising phoneme class, syllable stress, monosyllabicity of words, the duration of a phrase in words, etc;
- text structure (punctuation marks and conjunctions) also played a significant role in the modelling of segmental durations;
- syntactic, morphological and part-of-speech features of words affect the durations of segments in words; parts of speech yielded the best results in terms of interpretability;
- a comparison of different methods revealed that as far as the predictive precision is concerned, linear regression is an equal to the CART method and the neural networks method. In terms of interpretability, the best results were obtained with the CART method, the application of which, however, requires a phonetically balanced speech corpus.

The experience gained in the modelling of the temporal structure of speech enables us to maintain that statistical techniques based on speech corpora make it possible to predict segmental durations in a reliable way and to avoid major errors caused by a poor combination of rules. In addition, statistical methods can be used to discover and study small yet significant differences in temporal structure, such as the dependence of segmental durations on part of speech [P7].

A more precise modelling of segmental durations and pauses for TTS synthesis improves the quality of synthetic speech and enables us to automatically generate different voice profiles for the synthesis based on speech corpora.

What could have been done differently? Rather than recording a large number of speakers it would have been more useful to collect a bigger set of data per speaker. Perhaps we should have also confined ourselves to only one text type (e.g. news). In the phonetic database BABEL the passages read by each speaker were relatively short and therefore it was not possible to generate a pause model for these speakers. In some cases there was also too little speech material for the model of segmental durations. We should have also included in the list of features information on syllable structure, because the quantity degree of the foot is best realised in the portion starting with the vowel of the stressed syllable and ending with the vowel of the unstressed syllable (Eek, Meister 1997; Ross, Lehiste 2001). This is pointed out in subsection 6.5.2 where it is written that not all foot-forming phones are equally important in identifying the quantity degree of the foot.

The above-mentioned problems should be taken into account in future research. In the autumn of 2006, the Estonian corpus-based TTS synthesis project (Mihkla et al 2007) was launched in the framework of the National Programme for Estonian Language Technology. The speech database of the corpus-based project already contains about 50 minutes of speech material per speaker. The speech corpus is based on phonetically "rich" texts that contain all diphones, frequent words and phrases, many word forms, numbers and dates (Piits et al 2007). It is a solid foundation for applying the methodology proposed in the present study to the modelling of the temporal structure of speech.

Articles [P6], [P7] also refer to the need to carry out perception tests. As the main users of the TTS synthesiser are the blind and visually impaired, tests are being carried out in cooperation with the members of the North-Estonian Association for the Blind.

Another important direction in future research is corpus-based statistical modelling of other essential aspects of prosody – fundamental frequency and intensity. Some aspects of the modelling of fundamental frequency have already been touched upon in two articles referred to in this dissertation: modelling the intonation of questions with *kas*-particle [P1] and the relation between intonation and syntactic, morphological and part-of-speech features [P4]. Also, the speech melody of one radio announcer has been modelled by applying durational features. As the fundamental frequency and speech signal intensity are dependent on features that are, to a certain extent, different from duration, modelling experiments should be carried out to select significant features.

KIRJANDUS

Asu, Eva Liina 2004. The phonetics and phonology of Estonian intonation. PhD Thesis, University of Cambridge.

Breiman, L., Friedman, J., Olshen, R., Stone, C. 1984. Classification and regression trees. Monterey, CA, Wadsworth & Brooks.

Brinckmann, C., Trouvain, J. 2003. The role of duration models and symbolic representation for timing in synthetic speech. International Journal of Speech Technology 6: 21–31.

Campbell, Nick 2000. Timing in speech: a multilevel process. – Prosody: theory and experiment, M. Horne (editor). pp. 281–334, Dordrecht/Boston/London: 281–334, Kluwer Academic Publishers.

Campbell, N. W., Isard, S. D. 1991. Segment durations in a syllable frame. Journal of Phonetics 19: 37–47.

Clark, R., Richmond, K., King, S. 2007. Multisyn: Open-domain unit selection for the Festival speech synthesis system. Speech Communication, 49: 317–320.

Dutoit , Thierry 1997. An Introduction to Text-to-Speech Synthesis. Kluwer Academic Publishers, Dordrecht.

Eek, Arvo 1974. Observations on the duration of some word structures: I. Estonian Papers in Phonetics, EPP 1974:18–31.

Eek, Arvo 1987. The perception of word stress: a comparision of Estonian and Russian. – In honor of Ilse Lehiste (eds R. Channon, L. Shockey). Netherlands Phonetic Arhives VI: 19–32. Dordrecht (Holland), Providence (USA): Foris Publications.

Eek, Arvo, Meister, Einar 1997. Simple perception experiments on Estonian word prosody: foot structure vs segmental quantity. In: Lehiste, I.; Ross, J. (eds.). Estonian Prosody: Papers from a Symposium, Proceedings of the International Symposium on Estonian Prosody, Tallinn, Estonia, October 29–30, 1996. Institute of the Estonian Language and Authors, Tallinn: 71–99.

Eek, A., Meister E. 1999. Estonian speech in the BABEL multi-language database: Phonetic-phonological problems revealed in the text corpus. – *Proceedings of LP'98. Vol II, ed. O. Fujimura et al.*, Prague: 529–546, The Karolinum Press

Eek, Arvo and Meister, Einar 2003. Foneetilisi katseid ja arutlusi kvantiteedi alalt (I) : Häälikukestusi muutvad kontekstid ja välde. Keel ja Kirjandus, 46, 11: 815 – 837 & 12: 904–918.

Eek, Arvo and Meister, Einar 2004. Foneetilisi katseid ja arutlusi kvantiteedi alalt (II) : Takt, silp ja välde. Keel ja Kirjandus, 47, 4: 251–277 & 5: 336 – 357.

Fishel, Mark; Mihkla, Meelis 2006. Modelling the temporal structure of newsreaders' speech on neural networks for Estonian text-to-speech synthesis. - Proceedings of the 11th International Conference "Speech and Computer": SPECOM2006. St. Petersburg: Anatolya Publishers: 303–306.

Gurney, Kevin 1997. An introduction to neural networks. London, UCL Press.

Hint, Mati 1997. The Estonian quantity degrees in prosody and morphophonology. – Estonian Prosody: Papers from a Symposium, Proceedings of the International Symposium on Estonian Prosody, Lehiste, I.,; Ross, J. (eds.), Tallinn, Estonia, October 29–30, 1996. Institute of the Estonian Language and Authors, Tallinn: 125–135.

Hint, Mati 1998. Häälikutest sõnadeni. Tallinn, Eesti Keele Sihtasutus.

Holmes, J. N. 1988. Speech synthesis and recognition. Van Nostrand Reinhold. London.

Huggins, A. W. F. 1968. The perception of timing in natural speech: compensation within syllable. *Language and Speech* 11: 1–11.
Horák, Pavel 2005. Using neural networks to model Czech text-to-speech synthesis. – Proceedings of the 16th Conference of electronic speech signal processing, R. Vich (editor). pp. 76–83, Prague: 76–83, TUDpress.
Hosmer, D.W., Lemeshow, S. 2000. Applied logistic regression. New York, John Wiley & Sons.
Kaalep, Heiki-Jaan and Vaino, Tarmo 2001. Complete morphological analysis in the linguist's toolbox. – Congressus Nonus Internationalis Fenno-Ugristarum, Tartu 7.–13.08.2000: Tartu: TÜ Kirjastus, 2001, (V): 9–16.
Kaiki, N., Takeda, K., Sagisaka, Y. 1992. Linguistic properties in the control of segmental durations for speech synthesis. – Talking machines (eds G. Bailly, C. Benôit). Amsterdam: North-Holland: 255–264.
Keller, Eric 2007. Waves, beats and expectancy. – Proceedings of the 16th International Congress of Phonetic Sciences (eds. Jürgen Trouvain, William J. Barry). Saarbrücken, 6–10 August 2007. Saarbrücken: 355–360.
Keller, Eric, Port, Robert 2007. Speech timing: approaches to speech rhythm. – Proceedings of the 16th International Congress of Phonetic Sciences (eds. Jürgen Trouvain, William J. Barry). Saarbrücken, 6–10 August 2007. Saarbrücken: 327–329.
Klabbers, Esther 2000. Segmental and Prosodic Improvements to Speech Generation. PhD Thesis, Eindhoven University of Technology (TUE)
Klatt, D. H. 1979. Synthesis by rule of segmental durations in English sentences. – Frontiers of Speech Communication research, B. Lindblom & S. Öhman (eds.). New York: 287–300, Academic Press.
Klatt, D. H. 1980. Software for a cascade/parallel formant synthesizer. The Journal of the Acoustical Society of America, vol. 67: 971–995.
Krishna, N. S., Murthy, H. A. 2004. Duration Modelling of Indian Languages Hindi and Telugu. – Proceedings of 5th ISCA Speech Synthesis workshop, June 14–16, 2004. Carnegie Mellon University, Pittsburgh: 197–202.
Krull, Diana 1991. Stability in some Estonian duration relations. Institute of Linguistics, University of Stockholm, Perilius 13: 57–60.
Krull Diana 1992. Temporal and tonal correlates to quantity in Estonian. Phonetic Experimental Research, Institute of Linguistics,University of Stockholm (PERILUS) XV:17–36.
Krull, D., 1997. Prepausal lengthening in Estonian: Evidence from Conversational speech. – Estonian Prosody: Papers from a Symposium, Proceedings of the International Symposium on Estonian Prosody, Lehiste, I.; Ross, J. (eds.), Tallinn, Estonia, October 29–30, 1996. Institute of the Estonian Language and Authors, Tallinn: 136–148.
Lehiste, Ilse 1960. Segmental and syllabic quantity in Estonian. American Studies in Uralic Linguistics I, Bloomington, Ind, Indiana University: 21–82.
Lehiste, I. 1977. Isochrony reconsidered, Journal of Phonetics, vol. 5, 253–263.
Lehiste, I., 1981. Sentence and paragraph boundaries in Estonian. – *Congressus Quintus Internationalis Fenno-Ugristarum, Turku, 20.–27. 1980, Pars VI,* 1981: 164–169.
Lehiste, I., Fox, R. 1993. Influence of duration and amplitude on the perception of prominence by Swedish listeners. *Speech Communication* 13: 149–154.

Lehiste, Ilse 1997. Search for phonetic correlates in Estonian Prosody. – Estonian Prosody: Papers from a Symposium, Proceedings of the International Symposium on Estonian Prosody, Lehiste, I.; Ross, J. (eds.). Tallinn, Estonia, October 29–30, 1996. Institute of the Estonian Language and Authors, Tallinn: 11–35.

Liberman, A. M. 1959. Minimal rules for synthesizing speech. The Journal of the Acoustical Society of America: 1490–1499.

Liiv, Georg 1961. Eesti keele kolme vältusastme vokaalide kestus ja meloodiatüübid. Keel ja Kirjandus, 4, 7: 412–424 & 8: 480–490.

Liiv, G., Remmel, M. 1975. Estimate of the distinctive parameters in the domain of timing, fundamental frequency and intensity with implications for modelling of a quantitative system. – Proceedings of the Speech Communication Seminar, Stockholm August 1–3, 1974, vol. 2. Speech Production and Synthesis by Rules (ed G. Fant): 179–185. New York, London, Sydney, Toronto.

Lingaard, R. 1985. Electronic synthesis of speech. Cambridge University Press, Cambridge: 1–17.

Markel J. D., Gray A. H. 1976. Linear Prediction of Speech. Berlin/Heidelberg/New-York: Springer-Verlag.

Meister, E. 1991. Intonation modelling: A "contour interaction" based algorithm. – Papers from the 16th meeting of Finnish phoneticians, Oulu, Finland:, 1991: 69–74.

Meister, Einar and Werner, Stefan 2006. Intrinsic microprosodic variations in Estonian and Finnish: acoustic analysis. – Fonetiikan Päivät 2006 = The Phonetics Symposium 2006: (Toim.) Aulanko, R.; Wahlberg, L; Vainio, M.. Helsinki: University of Helsinki, 2006, (Publications of the Department of Speech Sciences, University of Helsinki): 103–112.

Mihkla, M.; Meister, E. 2002. Eesti keele tekst-kõne-süntees. Keel ja Kirjandus, 45(2): 88 – 97 ja 45(3): 173–182.

Mihkla, M.; Meister, E.; Eek, A. 2000. Eesti keele tekst-kõne süntees: grafeem-foneem teisendus ja prosoodia modelleerimine. Hennoste,T. (Toim.). Arvutuslingvistikalt inimesele. Tartu: 309–320.Tartu Ülikool.

Mihkla, Meelis, Piits, Liisi, Nurk, Tõnis, Kiissel, Indrek 2007. Development of a unit selection TTS system for Estonian. – Proceedings of the Third Baltic Conference in Human Language Technologies. Kaunas, Lithuania. October 4–5 2007 , Ilmumas.

Möbius, B., van Santen, J. 1996. Modeling Segmental Duration in German Text-to-Speech Synthesis. ICSLP 96: 2395–2398.

Piits, Liisi; Mihkla, Meelis; Nurk, Tõnis; Kiissel, Indrek 2007. Designing a speech corpus for Estonian unit selection synthesis. – Nodalida 2007 Proceedings: The 16th Nordic Conference of Computational Linguistics: 367–371.

Preminger, Alex, Brogan, Terry 1993. The New Princeton Encyclopedia of Poetry and Phonetics. Princeton, Princeton University Press.

Riley, Michael 1992. Tree-based modelling of segmental durations. – Talking machines (eds G. Bailly, C. Benôit). Amsterdam: North-Holland: 265–273.

Ross, Jaan, Lehiste, Ilse 2001. The temporal structure of Estonian runic songs. Berlin-New York, Mouton de Gruyter.

Sagisaka, Yoshinori 2003. Modeling and perception of temporal characteristics in speech. – Proceedings of 15th International Congress of Phonetic Sciences, M. J. Sole, D. Recasens & J. Romero (eds.). Barcelona: pp. 1–6.

van Santen, Jan 1998. Timing. – Multilingual text-to-speech synthesis. The Bell Labs Approach, Sproat, R. (editor), Kluwer Academic Publishers: 115–140.
Siil, Imre 1991. Estonian prosody model for speech synthesis. – Proceedings of the XIIth International Congress of Phonetic Sciences. Aix-en-Provence: 510–513.
Stout, Rex 2003. *Deemoni surm.* CD-versioon (loeb Andres Ots). Tallinn: Elmatar.
Särg, Taeve 2005. Eesti keele prosoodia ning teksti ja viisi seosed regilaulus. Dissertationes Folkloristicae Universitatis Tartuensis. Tartu, Tartu Ülikooli Kirjastus.
Zellner, Brigitte 1994. Pauses and the temporal structure of speech. – Fundamentals of speech synthesis and speech recognition. Ed. E. Keller. Chichester: John Wiley: 41–62.
Tatham, Mark and Morton, Katherine 2005. Developments in Speech Synthesis. John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester.
Tseng, C. 2002. The prosodic status of breaks in running speech: examination and evaluation. – Proceedings of Speech Prosody 2002, Aix-en-Provence, France: 667–670.
Vainio, Martti 2001. Artificial neural network based prosody models for Finnish text-to-speech synthesis. Helsinki: University of Helsinki.
Viitso, Tiit-Rein 2003. Phonology, morphology and word formation. – M. Erelt (ed.) Estonian Language. Linguistica Uralica. Supplementary Series, vol. 1: 9–92.
Viks, Ülle 2000. Eesti keele avatud morfoloogiamudel. – Arvutuslingvistikalt inimesele (Ed. T. Hennoste). Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1. Tartu: 9–36.
Weisberg, S. 1985. Applied linear regression. New York, John Wiley & Sons.
Wenk, B. J., Wioland, F. 1982. Is French really syllable-timed? Journal of Phonetics,10: 193–216.
Wiik, Kalevi 1985. Regelsynthese zur Lautquantität im Estnischen. Ostseefinnische Untersuchungen. Ergebnisse eines Finnisch-Sowjetischen Symposions (toim. H. Leskinen). Helsinki, Suomalainen Kirjallisuuden Seura: 129–137.
Wiik, Kalevi 1991. Foneetika alused. Tartu.
Wiik, Kalevi 1991. On a third type of speech rhythm: foot timing. – Proceedings of the XIIth International Congress of Phonetic Sciences, Aix-en-Provence, August 19–24, 1991, Vol 3: 298–301.
Õim, Haldur 1976. Kas inimkeel on päritav? – Keel, mida me uurime. Koost. M. Mäger. Tallinn, Valgus: 158–161.

ARTIKLITE KOOPIAD

I

Mihkla, Meelis; Pajupuu, Hille; Kerge, Krista; Kuusik, Jüri 2004.
Prosody modelling for Estonian text-to-speech synthesis. –
The First Baltic Conference. Human Language Technologies,
The Baltic Perspective, April 21–22 2004.
Riga: 127–131.

Prosody modelling for Estonian text-to-speech synthesis[1]

Meelis Mihkla[1], Hille Pajupuu[1], Krista Kerge[2] and Jüri Kuusik[3]

[1] Institute of the Estonian Language, Tallinn, Estonia

[2] Tallinn University of the Educational Sciences, Tallinn, Estonia

[3] University of Tartu, Tartu, Estonia

E-mail: meelis.mihkla@eki.ee, hille.pajupuu@eki.ee, kristake@tpu.ee, jkuusik@deloittece.com

Abstract

An annoying and troublesome obstacle impeding the regular application of the text-to-speech synthesiser of the Estonian language developed in 1997–2002 has been the monotony of the output speech and its poor fluency. This article is concerned with the issues of modelling the prosody of synthesised speech. Under scrutiny is the steady improvement of intonation of interrogative sentences, and the link between pauses and prepausal lengthenings, and syntactic parsing of the text, instrumental to provide for natural rhythm of the output speech. Also presented is the predicting of durations of sounds by means of statistical models.

1. Introduction

Unlike the musical notation, there are no marks in the ordinary text, besides the punctuation that denote speech rate, pitch, intonation, pause, or stress. Interpretation of a written text is largely the matter of one's individual choice or judgement. The computer however has no power of free decision in this respect, because discretion calls for understanding of the content. Modelling of the prosodic structure of speech is a most complicated part of speech synthesis, involving as it does the generation of the duration of sounds and the melody contour (fundamental frequency contour of speech) corresponding to a sentence type.

In the earlier prosody model of the synthesis text-to-speech of Estonian (Mihkla et al., 2000) the values of durations used to base largely on "laboratory speech" (isolated words and sentences); intonation was modelled by the method of acoustic stylisation between linear declination curves.

The chapter on intonation of the question focuses on improvement of intonation of questions with *kas*-particle and that of special questions. On the basis of contour types, the respective rules of generation of intonation have been developed.

Speech contains complicated temporal patterns, which the text-to-speech system should be able to simulate, if the speech is to sound natural. More specifically, the system text-to-speech must be capable of generating the durations of sounds and pauses, not noticeably different from the values of the real speech. The chapter on temporal structure of speech concentrates on predicting of durations of sounds and pauses, on the basis of text, by use of statistical models.

2. Intonation of the question

Improvement of the intonation of sentence was started with intonation of the question, because that sounded the most unnatural.

Among all questions in written texts of Estonian, the general ones account for 53% and the special ones for 47%. 23% are general and general special questions formed with *kas*-particle. Modelling was started with that largest group.

The analysis of human speech showed that the interrogative sentences with *kas*-particle are clearly characterised by three fundamental frequency contour type: 1) with the general

[1] This work was completed within the framework of Grant no. 5039 of Estonian Science Foundation.

question contour, the prevailing tendency is to stress all longer words of the sentence; 2) with the general special question contour the questioner stresses the word at which he addresses the question; 3) with alternative question contour the questioner stresses both alternatives separated by *või ('or')*. The question word *kas* invariably turned out unstressed, the fundamental frequency contour only shot up thereafter, to decline by the end of the question. The question was posed 15% quicker, as compared with the declarative sentence. (Kerge et al., 2002; Mihkla et al., 2003).

For the synthesiser, out of three fundamental frequency contours only two were formulated as a rule, the contour of general special question and that of alternative question (Figures 2 and 3), because one group of the users of synthesiser – visually impaired and stark blind – did not wish the synthesiser to stress too many words in the sentence, like is the case with general question of *kas*-particle (which are tiring to an ear). The perception tests revealed that the intonation of *kas*-questions synthesised by rules created, was acceptable to listeners (Mihkla et al., 2003).

To form special questions, there are available in Estonian ca. 350 question words and -phrases. The analysis showed that unlike the *kas*-particle, the question word in a special question is always stressed. Hence, the contour of a special question swings up at the beginning, to drop by the end of the sentence (cf. Figure 3).

Like in the general special question formed with *kas*-particle (cf. Figure 1), the questioner in many cases stresses the word at which the question is targeted (cf. Figure 2). In case of special questions, however it is hard to identify the stressed word (focus) without making recourse to semantic means. Therefore, when formulating a rule for the synthesiser of the fundamental frequency for special question, only the question word was stressed, at first.

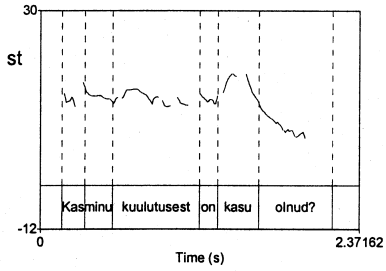


Figure 1. General special question "Kas minu kuulutusest on kasu olnud ?" *'Has there been any use of my advertisement?'* Stress falls on the word or phrase of which the question is asked. (Mihkla et al., 2003)

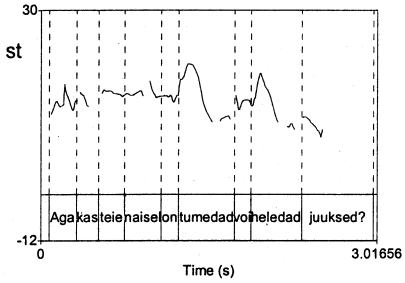


Figure 2. Alternative question "Aga kas teie naisel on tumedad või heledad juuksed?" *'By the way, does your wife have dark or light hair?'* The alternatives offered for an answer are stressed. (Mihkla et al., 2003)

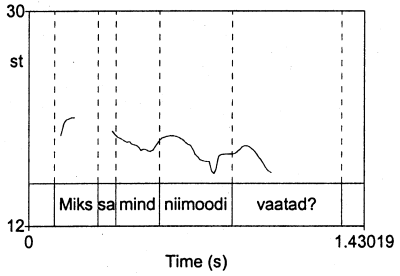


Figure 3. Special question "Miks sa mind niimoodi vaatad?" *'Why do you look at me like that?'* The question word '*Why*' is stressed

3. Modelling of the temporal structure of the speech

The existence of larger speech corpuses provides an opportunity to realise the transformation of text-prosody by means of statistical models.

3.1. Source material

Because we are concerned with text-to-speech synthesiser, the source material was a sample of texts read by announcers. On the basis of one-to-one correspondence of text and speech, it is possible to move over from symbols presentation of prosody to the acoustic one and also to establish whether and to what extent the syntactic parsing of the text is related to the prosodic parsing of the speech.

Taken as source material were passages of speech from the CD-version of a detective story read by an actor (Stout, 2003) and passages of speech and texts from the longer news read by announcers of Estonian Radio. Altogether, 12 speech passages were analysed, each 1-2 minutes long. All passages of speech were segmented into sounds and pauses.

3.2. Pauses and prepausal lengthenings

Prior to application of a general statistical model, pauses and prepausal lengthenings in speech were analysed, basing on material.

The first idea was to verify whether the pauses can be classified (for instance whether the pauses between phrases differ significantly from the pauses of the end of sentence or end of paragraph). Table 1 presents the average durations of pauses with different announcers, the generalised mean and standard deviation. It turns out that with a read text of normal speech rate the classification of speech pauses is perfectly possible and can be applied, in speech synthesis. Dispersion of values however is quite large. In speech recognition such classification is of little avail.

Table 1. Durations of pauses (ms) in speech, Σ - mean, σ –standard deviation

	Phrase end pauses	*Sentence end pauses*	*Paragraph end pauses*
Actor	270	558	1065
Male announcer	224	828	1008
Female announcer	197	769	1119
Generalised mean	Σ=229 σ=181	Σ=690 σ=261	Σ=1065 σ=265

Held in perspective in what follows was whether and how extensively the prosodic parsing of the speech is in correlation with syntactic parsing of the text, in cases when the latter is marked by punctuation and conjunctions.

Table 2. Connection of pauses with parsing of the text

	No of parsings in the text	*No of pauses in speech*	*%*
Paragrph end	21	21	100
Sentence end	58	56	97
Comma	80	41	51
Conjunction	22	7	32
Colon	7	7	100
Dash	14	13	93

As evidenced in Table 2, in speech a pause[2] is at the end of every paragraph and at the end of almost every sentence (only actor has defaulted on that rule and has merged the sentences). There is a very strong connection between syntax and prosody in case of colon and dash. Half the commas are related to pauses. The least marked in speech are phrases starting with such co-ordinating conjunctions, which do not require the comma. As it is, pauses are only one way of marking verbal parsing of the text. In the following stage, we tackled the connection of boundary lengthenings with pauses and syntactic parsing.

Taking a general look at foot lengthenings in the research material reveals that prepausal lengthenings do not provide the opportunity to classify, similar to that of pauses. The concept "prepausal lengthening" describes in the given material of Estonian only ca. 70% (71% pauses are preceded by word or foot lengthening). As evidenced in Lehiste's perception tests (Lehiste and Fox, 1993), the Estonians do prefer on the last syllable of the sentence a significantly lesser boundary lengthening as e.g. English speaking persons.

Table 3 presents the link between boundary lengthening and syntactic parsing of the text.

Table 3. Link of boundary lengthening with text parsing

	No of parsings in text	*No of foot lengthe-nings*	*%*
Paragraph end	21	15	71
Sentence end	58	39	67
Comma	80	42	53
Conjunction	22	12	55
Colon	7	4	57
Dash	14	13	93

[2] In this work, regarded as the prosodic pause has been the interruption of speech over 100 msec

Among the marks of punctuation, the lengthening is obviously related to dash. Apparently the connotation is suggested by the shape of the sign – the dragged line prompts the drawl. With pauses, formal rules can be applied solely on the basis of syntax. To model the prepausal lengthenings, however the unsophisticated approach is clearly not enough. Here, specific statistic models should be used and the lengthenings calculated as durations of the sounds, depending on the context.

3.3. Statistic modelling of durations of sounds

Serving as input data to statistical analysis was the sequence of sounds (phonemes) and durations of sounds, obtained by segmenting the speech wave. Basing on segmenting data and on the text corresponding to speech, a vector of features was composed, for every sound. Because our aim was to compare the suitability of different data processing methods (regression analysis, neural networks) for modelling of prosody, we selected the features so as to make them compatible with different methods.

We proceeded from the premise that every sound has its intrinsic duration, every sound belongs to a certain sound class (front vowels, plosive consonants, nasals etc.), the properties of which translate to the members of the given class; we also proceeded from the premise that neighboring sounds affect one another and the duration may be affected by structure of word and sentence.

Features were described on several hierarchical levels (level of phoneme, syllable, foot, word, phrase and sentence). Features of the phonemic level have been presented on Figure 4. When selecting the structure of phonemic features we followed the technique of the features' vector used by Martti Vainio in his neural networks experiments (Vainio, 2001). Unlike Vainio we presumed, however that every sound is directly impacted, from both sides, by one neighbouring sound only (in the aforementioned experiments the analysis involved 2–3 neighboring sounds). A phoneme is described, as a general case with the help of three features: identity, class and length. All sounds have been distributed into nine classes: vowels into three classes (high, mid, low), consonants into five classes (plosive, fricative, nasal, liquid, semivowel), plus a special class of pauses. In Estonian a sound has, phonologically two degrees of lengths – short and long (Eek and Meister, 2003).

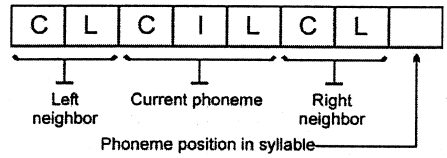


Figure 4. Data representation in phoneme-level (C-board class, L-length, I-identity)

To compare, the features of foot are quantity degree, length of foot in syllables and the situation of foot in word.

The output of model or the response variable – duration – is presented as logarithm LN (duration), because the logarithmed duration is more compliant with normal distribution.

Table 4. Summary of fit and the analysis of variance for the regression model

Summary of Fit					
Mean of Response	-2.5478		R-Square	0.4170	
Root MSE	0.3612		Adj R-Sq	0.4096	
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
Model	38	279.4	7.3545	56.37	<.0001
Error	2995	390.7	0.1305		
C Total	3033	670.2			

The initial results of statistical modelling of multiple regression analysis revealed that the model created is statistically significant (cf. Table 4). The analysis of regression coefficients disclosed that posing as significant features, for predicting the duration of the sound, were the length of the current sound (short or long), the class of the next sound, stress of the syllable, position of the syllable in foot, the length of the foot in syllables, and location of the word in phrase. Significance probability of other features was larger than significance level by 0.05, therefore it makes little sense to embrace them in regression model. Curiously the quantity degree of foot, being the cornerstone of Estonian word prosody, was not a significant feature for prediciting the duration of sound.

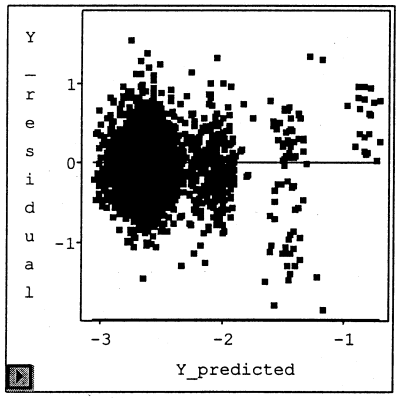


Figure 5. Distribution of prediction residuals

Those modelling results, however have been obtained relying on partial data volumes, only. As yet unresolved is the problem of multicollinearity, and outstanding is the analysis of outliers. Hence we will not jump at any rash conclusions, lest we commit a grave error.

The analysis of regression residues (errors) (cf. Figure 5) shows that catching the eye in error distribution are three "data clusters" dispersed with respect to each other. Error analysis calls for further details, however under a tentative surmise, the cause is the incorporation of pauses among the body of sounds data. It may well be that the pause is such a specific phenomenon that its duration should be modelled on the level of word, rather than sound.

4. Conclusions. Now what?

Described in this work have been the initial results and attempts, on the way to render more natural the prosody of the output speech of text-to-speech synthesiser of Estonian. Besides interrogative sentences, improvement of intonation is the order of the day, also with other sentence types. To predict the durations by data processing methods, the speech corpus should be replenished and various methods tried out. Predicting of the prosody data could be also applied, regarding the fundamental frequency and intensity of signal.

5. Bibliographical References

Arvo Eek and Einar Meister. 2003. Foneetilisi katseid ja arutlusi kvantiteedi alalt (I). *Keel ja Kirjandus*, 11, pp. 815-837.

Krista Kerge, Meelis Mihkla and Hille Pajupuu. 2002. Modelling.the Estonian general questions with the *kas*-particle. In P. Korhonen, ed., *Fonetiikan Päivät 2002. The Phonetics Symposium 2002. Helsinki University of Technology Laboratory of Acoustics and Audio Signal Processing.* Report 67. Espoo, pp. 187-194.

Ilse Lehiste and Robert Fox. 1993. Influence of duration and amplitude on the perception of prominence by Swedish listeners. *Speech Communication,* 13, pp. 149-154.

Meelis Mihkla, Einar Meister and Arvo Eek. 2000. Eesti keele tekst-kõne süntees: grafeem-foneem teisendus ja prosoodia modelleerimine. *Arvutuslingvistikalt inimesele. TÜ üldkeeleteaduse õppetooli toimetised*, 1, Tartu, pp. 309-320.

Meelis Mihkla, Hille Pajupuu and Krista Kerge. 2003. Modelling and perception of the Estonian general questions with the *kas*-particle. *Proceedings of 15th International Congress of Phonetic Sciences.* Barcelona, pp. 539-542.

Rex Stout. 2003. *Deemoni surm.* CD-versioon (loeb Andres Ots). Elmatar, Tallinn.

Martti Vainio. 2001. *Artifical neural network based prosody models for Finnish text-to-speech synthesis.* University of Helsinki, Helsinki.

II

Mihkla, Meelis; Kuusik, Jüri 2005.
Analysis and modelling of temporal characteristics of speech
for Estonian text-to-speech synthesis. Linguistica Uralica, XLI(2): 91–97.

MEELIS MIHKLA, JÜRI KUUSIK (Tallinn)

ANALYSIS AND MODELLING OF TEMPORAL CHARACTERISTICS OF SPEECH FOR ESTONIAN TEXT-TO-SPEECH SYNTHESIS*

Abstract. A text-to-speech system must be capable of generating sounds and pauses with such durations that do not noticeably differ from natural speech. Currently, the prosodic modelling of Estonian text-to-speech synthesis is largely based on generalized measurements of speech units in isolated words and sentences, and as a result the synthesized speech is often monotonous and has poor fluency. In this work the first attempts are made to improve the naturalness of the output speech of the speech synthesiser with the help of statistical duration models of fluent speech. The source material consisted of (a) prose read out by a professional actor, and (b) news broadcasts read by announcers. On the basis of this material variability of the duration of pauses and boundary lengthenings was investigated. It turns out that in the case of a read text at normal speech rate the classification of speech pauses is perfectly possible and can be applied in speech synthesis. An attempt was also made to establish whether and to what extent the syntactic parsing of a text is related to the prosodic parsing of speech. A generalized regression analysis revealed what features are essential in predicting sound durations in speech and a statistically optimal model was developed. Curiously the quantity degree of a foot, despite being the cornerstone of Estonian word prosody, was not a significant feature for prediciting the duration of a sound on the basis of this material. The results of the modelling were then compared with the expert opinions of some Estonian phoneticians.

Keywords: Estonian, duration of sounds, pause, statistical modelling, regression analysis, text-to-speech synthesis.

1. Introduction

The task of text-to-speech synthesis is to convert orthographic text into natural-sounding speech. For the artificial speech to sound realistic to the human ear, it should comprise realistic intonation, rhythm and stress patterns. More specifically, the text-to-speech system must be able to generate durations of sounds and pauses not notably different from the values of the actual speech.

Currently, prosodic modelling in Estonian text-to-speech synthesis (Mihkla, Meister, Eek 2000) is largely based on generalized measurements

* Support from the Estonian Science Foundation, grant No. 5039, and state program "Estonian language and national memory" has made the present work possible.

of speech units in isolated words and sentences. The resulting output (synthesized) speech, however, is often monotonous and has poor fluency, which sets application limitations on the synthesizer. As indicated by Nick Campbell, the durations of sounds in isolated words or sentences are largely different from durations of sound in the fluent speech (Campbell 2000). The speech contains complicated temporal patterns, which the text-to-speech system must be able to imitate for the speech to sound natural. The availability of oral speech corpuses provides an opportunity to achieve the text-prosody transformation with the help of statistical models.

In this work the first attempts are made to improve the naturalness of the output speech of an Estonian speech synthesiser with the help of statistical duration models of fluent speech. We applied the technology of regression analysis to find out the essential features of sound durations and to compose a prediction model. The results of modelling the durations are compared with expert opinions given by Estonian phoneticians. With the aim to providing for a natural rhythm of the output speech, the relation of pauses and boundary lengthenings with syntactic parsing of the text is studied.

2. Source material

Because we are concerned with a text-to-speech synthesiser, the source material was a sample of texts read by announcers. On the basis of one-to-one correspondence of text and speech, it is possible to move from a symbol-based representation of prosody to the acoustic one and also to establish whether and to what extent the syntactic parsing of the text is related to the prosodic parsing of the speech.

The source material consisted of passages of speech from the CD-version of a detective story read by an actor (Stout 2003) and passages of speech and texts from longer news read by announcers of Estonian Radio. Altogether, 12 speech passages were analysed, each 1-2 minutes long. All passages of speech were segmented into sounds and pauses.

3. Analysis of pauses and boundary lengthenings

Prior to the application of a general statistical model, pauses and prepausal lengthenings in speech were analysed, based on this material. The pauses and prepausal lengthenings in Estonian speech have been studied cursorily or intermittently, as a by-product in the context of other tasks. Ilse Lehiste (1981) verified whether prepausal lengthenings were in correlation with the length of subsequent pauses and she established an extremely weak link. Diana Krull (1997) studied prepausal lengthenings in dialogue in two-syllable words in the context of quantity degree. Arvo Eek and Einar Meister (2003) looked at end-of-sentence lengthenings on the basis of tempocorpus. However, they examined only words of a specific structure, and focused on quantity degree features. Therefore the need became evident to measure, for Estonian language text-to-speech synthesis, pauses and boundary foot lengthenings, on the basis of a text read out from real speech.

With a view to analysing the pauses and foot lengthenings, the durations of pauses derived from the speech wave were measured, and the foot lengthenings were calculated. For the calculation of foot lengthenings, the durations

of sounds comprising the foot were summed, after which the actual duration was compared to the mean duration of the given foot structure in the speech of that announcer. The first hypothesis was to verify whether pauses and foot lengthenings could be classified (for instance, whether the pauses between phrases[1] differ significantly from the sentence end or paragraph end pauses).

Table 1

Durations of pauses and boundary lengthenings (ms) in speech

Dictors	Phrase end pauses	Sentence end pauses	Paragraph end pauses	Phrase end lengthenings	Sentence end lengthenings	Paragraph end lengthenings
Actor1 (m)	352	558	1025	200	220	315
Announcer1 (f)	303	828	902	124	112	117
Announcer2 (m)	286	769	1132	95	90	122
Generalised mean	**323**	**678**	**1021**	**155**	**161**	**217**

Table 1 presents the mean durations of pauses as per announcers and the generalised mean. Looking at the generalised means suggests that in case of a text read out at normal speech rate the classification of speech pauses is fully possible. The statistical analysis of samples corroborates this surmise. Analysis of pairs of the logarithmic durations of pauses with the help of a Student t-test reveals that the t-statistic values on significance level $p = 0.01$ noticeably exceed the t-critical two-tail quantile (cf. Table 2) on probability of significance of hypothesis $p < 0.0001$. Hence it seems proved that the mean values of durations of pauses differ and the classification of pauses is fully possible, which fact could be applied in speech synthesis. The dispersion and variance however are large; therefore in speech recognition, for instance, such classification is to no avail.

When analysing, with the help of Student t-test the data of foot lengthenings (cf. Table 2) we had to accept the null hypothesis: the foot lengthenings are from samples of the same mean value.

Table 2

Student t-test results for comparison of pairs of sample means (Ph-Se — between phrase and sentence, Ph-Pa — between phrase and paragraph, Se-Pa — between sentence and paragraph)

	Pauses			Foot lengthenings		
	Ph-Se	**Ph-Pa**	**Se-Pa**	**Ph-Se**	**Ph-Pa**	**Se-Pa**
T stat	8.87	12.25	5.91	0.81	0.26	0.65
T critical two-tail	2.62	2.76	2.72	2.65	2.84	2.90
P (T <= t)	< 0.0001	< 0.0001	< 0.0001	0.42	0.79	0.52

Next examined was whether and to what extent the prosodic parsing of speech correlates with syntactic parsing where the latter is indicated by punctuation marks and conjunctions. As shown in Table 3, there is invariably a pause in speech[2] at the paragraph end and the sentence end. In case of a colon and dash too there is a strong correlation between syntax

[1] In this work, the phrase means the clause or element of enumeration, which has been determinated within the sentence by punctuation mark or conjunction.

[2] In this work we have treated as a prosodic pause an interruption of speech over 50 ms.

and prosody. Half the commas are related to pauses. The least marked in speech are phrases starting with those co-ordinating conjunctions which do not require the comma.

Table 3

Connection of pauses and foot lengthenings with the text parsing

	No. of parsings in the text	No. of corresponding pauses in the speech		No. of corresponding foot lengthenings in the speech	
		Cnt	%	Cnt	%
Paragraph end	21	21	100	15	71
Sentence end	58	58	100	39	67
Comma	80	41	51	42	53
Conjunction	22	7	32	12	55
Colon	7	7	100	4	57
Dash	14	13	93	13	93

Among the punctuation marks, lengthening is obviously related to the dash. Apparently the connotation is suggested by the shape of the sign — the stretched line prompts the drawl. Suggestive of the link between pauses and boundary lengthenings is the English term 'prepausal lengthening'. This term applies, on the basis of this Estonian language speech material, only 70% of the time (only 143 pauses are preceded by word or foot lengthening). According to perception tests carried out by I. Lehiste (Lehiste, Fox 1993) Estonian speakers expect significantly less final lengthening on the last syllable of the sentence than English speakers do.

But if we wish to lend synthetic speech a natural rhythm, it does not suffice if we just find out the mean durations of pauses and lengthenings. Instead, we should rather model their durations and temporal positions in a context-sensitive way.

4. Statistical modelling of segmental durations

Because we are still seeking the most suitable statistical method to predict the durations, we carried out regression analysis on the basis of partial source material (passages from the detective story read by the actor). The input data to the statistical analysis of durations was the sequence of sounds (phonemes) and sound durations obtained by segmenting the speech wave. On the basis of a text corresponding to the speech, we formed a vector of features (with 17 features) for every sound. Those argument features were described on several hierarchical levels (phoneme, syllable, foot, word, phrase and sentence levels). We proceeded from the presumption that every sound has intrinsic duration, a vowel belongs to a concrete class of sounds (front vowels, plosive consonants, nasals etc) whose properties translate to the members of the given class; and also, that adjacent sounds impact on one another and that the duration may be influenced by both the word and the sentence structure. The output of the model or the functional feature (response) — duration — was presented as logarithmic LN (duration), because the logarithmic duration conforms more to the normal distribution. Because the argument features (explanatory variables) were numerous, an optimum selection had to be made among them, i.e. we had to locate the features most likely to affect the response.

Table 4

Summary of fit and the analysis of variance for the regression model of durations

Summary of Fit

Mean of Response	–2.7530	R-Square	0.5393
Root MSE	0.2886	Adj R-Sq	0.5368

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
Model	26	478.5	18.403	220.87	<0.0001
Error	4906	408.8	0.0862		
C Total	4932	887.2			

The initial results of statistical modelling of multiple regression analysis revealed that the model created is statistically significant (cf. Table 4). The analysis of regression coefficients disclosed that significant features for predicting the duration of the sound were the class and length (short or long) of the current sound, the class of the next sound, the position of the sound in syllable, the position of the syllable in foot, the length of the word in feet, and the location of the word in phrase. Curiously the quantity degree of the foot, despite being the cornerstone of Estonian word prosody, was not a significant feature for prediciting the duration of a sound. Those modelling results, however, have been obtained relying on only partial data volumes. Table 5 presents the features estimated as significant by experts and the statistically significant argument features obtained by regression analysis. Acting as experts were six Estonian phoneticians. The conclusions of the experts and the results of regression analysis coincided on average to 49%.

Table 5

Expert opinions versus results of regression analysis
(ExpN — N expert, Reg — results of regression analysis,
1 — significant explanatory variable, 0 — unsignificant variable)

Explanatory variable	Exp1	Exp2	Exp3	Exp4	Exp5	Exp6	Reg
Previous phoneme class	0	0	0	0	0	0	0
Previous phoneme length	1	1	1	1	0	0	0
Current phoneme class	1	1	1	0	1	0	1
Current phoneme length	1	1	1	1	0	0	1
Next phoneme class	1	1	0	0	0	0	1
Next phoneme length	1	0	1	1	0	0	0
Phoneme position in syllable	1	1	0	1	0	0	1
Stress of syllable	1	1	1	1	1	1	0
Type of syllable	1	0	0	1	1	1	0
Quantity degree of foot	1	1	1	1	0	1	0
Syllable position in foot	1	1	1	1	0	1	1
Length of foot in syllables	1	1	0	1	0	1	0
Foot position in word	1	0	0	1	0	1	0
Length of word in feet	1	1	0	1	0	0	1
Word position in phrase	1	1	1	1	0	0	0
Length of phrase in words	1	0	0	1	0	1	1
Length of sentence in phrases	1	0	0	0	0	0	0
Total "correct" answers	8	11	8	7	9	7	
%	47%	65%	47%	41%	53%	41%	

Total average 49%

The analysis of prediction residuals or errors (cf. Figure 1) showed that in the distribution of errors there were three "data clusters" distanced from one another. A closer look revealed that the two right-hand clusters were constituted by pauses. The residuals may be considered, at a visual estimate, to be homoscedastic.

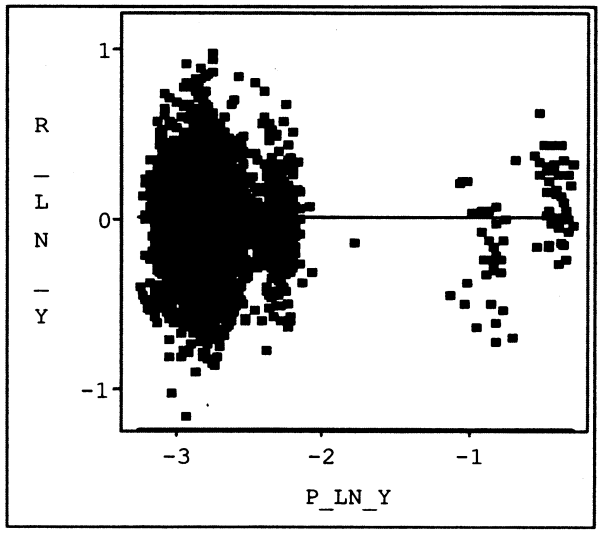


Figure 1. **Residual by predicted values of sound durations: y-axis (R_LN_Y) — prediction residuals of logarithmic sound durations, x-axis (P_LN_Y) — predicted values of logarithmic sound durations.**

5. Conclusions and future work

This paper has described the preliminary results and the first attempts to make the prosody of the output speech of a text-to-speech synthesiser of Estonian more natural. The analysis of prediction errors showed that the sounds and pauses should be handled separately at analysis. To predict the duration of sounds and pauses using statistical methods the volume of material analysed should be expanded, with various methods tested (e.g. neural networks).

REFERENCES

C a m p e l l, N. 2000, Timing in Speech. A Multilevel Process. — Prosody. Theory and Experiment, Dordrecht—Boston—London, 281—334.

E e k, A., M e i s t e r, E. 2003, Foneetilisi katseid ja arutlusi kvantiteedi alalt (I). Häälikukestusi muutvad kontekstid ja välde. — KK, 815—837.

K r u l l, D. 1997, Prepausal Lengthening in Estonian: Evidence from Conversational Speech. — Estonian Prosody: Papers from a Symposium. Proceedings of the International Symposium on Estonian Prosody, Tallinn, Estonia, October 29-30, 1996, Tallinn, 136—148.

L e h i s t e, I. 1981, Sentence and Paragraph Boundaries in Estonian. — CIFU V, Pars VI, 164—169.

Lehiste, I., Fox, R. 1993, Influence of Duration and Amplitude on the Perception of Prominence by Swedish Listeners. — Speech Communication 13, 149—154.

Mihkla, M., Meister, E., Eek A. 2000, Eesti keele tekst-kõne süntees: grafeem-foneem teisendus ja prosoodia modelleerimine. — Arvutuslingvistikalt inimesele, Tartu (Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1), 309—320.

Stout, R. 2003, Deemoni surm. CD-versioon. Loeb Andres Ots, Tallinn.

МЕЕЛИС МИХКЛА, ЮРИ КУУСИК (Таллинн)

АНАЛИЗ И МОДЕЛИРОВАНИЕ ВРЕМЕННЫХ ХАРАКТЕРИСТИК РЕЧИ ДЛЯ ЭСТОНСКОГО ТЕКСТ-РЕЧЬ-СИНТЕЗА

Текст-речь-система должна быть способной генерировать звуки и паузы продолжительностью, которая не отличалась бы значительно от подобных характеристик в обычной речи. Моделирование настоящей просодии эстонского текст-речь-синтеза базируется в основном на обобщенных результатах измерений речевых единиц изолированных слов и предложений. Поскольку выходная речь синтезатора зачастую оказывается монотонной, а слитность речи неустойчивой, широкое использование синтезатора ограничено. В данной работе сделаны первые попытки улучшить естественность выходной речи, используя для этого статистические модели слитной речи. В качестве исходного материала служили подчитка беллетристического текста профессиональным актером и чтение новостей диктором. По этим материалам исследовалась вариативность длительности пауз и предпаузное удлинение. Оказывается, если текст подчитан в нормальном темпе, то классификацию пауз вполне можно использовать в синтезе речи. Была предпринята также попытка выяснить, насколько синтаксическое членение текста связано с просодическим членением речи. С помощью регрессионного анализа выявлялись признаки, важные при прогнозировании длительности звуков, и строилась статистически оптимальная модель. Удивительно, что степень количества речевого такта, которая в эстонском языке является краеугольным камнем просодии слова, отнюдь не служит важным параметром прогнозирования длительности звуков в данном материале. Результаты моделирования сравнивались с экспертной оценкой.

III

Mihkla, Meelis 2005. Modelling pauses and boundary lengthenings in synthetic speech. – Proceedings of the Second Baltic Conference on Human Language Technologies, April 4–5, 2005. Tallinn: 305–310.

MODELLING PAUSES AND BOUNDARY LENGTHENINGS IN SYNTHETIC SPEECH

Meelis Mihkla
Institute of the Estonian Language, Tallinn, Estonia

Abstract

In order to make synthetic speech, generated from a given text, sound natural, one has to exert strict control over the temporal structure of the speech flow. At that, close attention should be paid to pauses and boundary lengthenings as phrase markers and vital factors of speech rhythm in general. The present study analyses the duration of pauses and boundary lengthenings in various Estonian texts (fiction, news, other) read out by 27 dictors. If we wish to lend synthetic speech a natural rhythm, it does not suffice if we just find out the mean durations of pauses and lengthenings. Instead, we should rather model their durations and temporal positions in a context-sensitive way. In this study the duration of pauses and boundary lengthenings as well as their temporal positions in synthesised speech have been modelled on the basis of text structure, using some basic methods of prediction (regression analysis). [1]

Keywords: pause, boundary lengthening, synthetic speech, regression analysis

1. Introduction

For the artificial speech to sound realistic in human ear, it should comprise natural-sounding intonation, rhythm and stress pattern. The temporal phenomena like pauses and syllable lengthenings constitute a vital part of prosodic aspects of speech. The pre-boundary lengthenings of intonation phrases are often applied in the synthesising devices, however pause modelling has less frequent currency. The reason might be that the pauses are largely variable both in duration and in location in speech flow. The duration of pauses and boundary lengthenings and their situation in speech flow depends on sentence structure, speaker and also the given language (Zvonik, Cummins 2002).

The pauses and prepausal lengthenings in the Estonian language speech have been studied cursorily or intermittently, as the by-product in the context of other tasks. Ilse Lehiste verified (Lehiste 1981) whether prepausal lengthenings were in correlation with the length of subsequent pauses and she established an extremely weak link. Diana Krull studied prepausal lengthenings in dialogue speech in two-syllable words in the

[1] This work was completed within the framework of Grant no. 5039 of Estonian Science Foundation and the state program „Estonian language and national memory".

context of quantity degree (Krull 1997). Arvo Eek and Einar Meister looked into the end-of-sentence lengthenings on the basis of tempocorpus (Eek, Meister 2003). However, they held under scrutiny only the words of a specific structure, and focused on quantity degree features.

Therefore a need evolved, to measure for the Estonian language text-speech synthesis, the pauses and the boundary foot lengthenings, on the basis of a text read out from real speech, and to model their durations and locations in the speech flow.

2. Source material

Because we are concerned with text-to-speech synthesiser, the source material was a sample of texts read by announcers. Under condition that there is one-to-one conformity of text and speech, the symbol representation of prosody may be replaced by an acoustic representation, whereas it is possible to establish whether and for what extent the syntactic parsing of text is related to the prosodic parsing of speech.

Elected as the base material were:

- Passages of speech from the CD-version of a detective story (Stout 2003) read by an actor;
- Passages of speech and texts from the longer news from Estonian Radio, read by announcers;
- Passages of speech from the Estonian phonetic database BABEL (Eek, Meister 1998).

Altogether, 44 passages of speech were analysed (each 0.4-2 minutes long), in the presentation of 27 speakers (14 men and 13 women). All passages of speech were segmented into sounds and pauses.

3. Analysis of pauses and prepausal lengthenings

With a view to analysing the pauses and foot lengthenings, the durations of pauses derived from the speech wave were measured, and the foot lengthenings were calculated. For calculation of foot lengthenings, the durations of sounds comprising the foot were summed up, after which the actual duration was compared to the mean duration of the given foot structure in the speech of a concrete announcer. The first hypothesis was whether pauses and foot lengthenings could be classified (for instance, whether the pauses between phrases[2] differ significantly from the sentence end or paragraph end pauses). Presented in Table 1 are the mean durations of pauses as per announcers, the mean values for male and female announcers and the generalised mean. As is seen, the variance of even the mean values is very large. Curiously however, the generalised means of male and female pauses differ from one another as per durations within 10% only. The general mean visual observations suggest that in case of a text read out at normal speech rate the classification of speech pauses is fully possible. The statistical analysis of samples corroborates this surmise. The analysis in pairs of the logarithmic durations of pauses with the help of Student t-test reveals that the t-statistic values on significance level p=0,01 noticeably exceed the t-critical two-tail quantile (cf. Table 2). Hence it seems proved that the mean values of durations of pauses differ and the classification of pauses is fully possible, which fact could be applied in speech synthesis, for that matter.

[2] In this work, the phrase means the clause or element of enumeration, which has been determinated within the sentence by punctuation mark or conjunction.

Table 1. Durations of pauses and boundary lengthenings (ms) in speech

Dictors	*Phrase end pauses*	*Sentence end pauses*	*Paragraph end pauses*	*Phrase end lengthenings*	*Sentence end lenthenings*	*Paragraph end lengthenings*
Actor1 (m)	352	558	1025	200	220	315
Announcer1 (f)	303	828	902	124	112	117
Announcer2 (m)	286	769	1132	95	90	122
Speaker1 (f)		547		103	107	113
Speaker2 (m)	361	862		60	73	77
Speaker3 (f)	255	306		76	138	
Speaker4 (m)	145	478		78		89
Speaker5 (f)	275	879		109	100	88
Speaker6 (f)	470	1179		89	89	
Speaker7 (f)	117	559		85		64
Speaker8 (m)	416	829		73	118	75
Speaker9 (m)	260	683		94	91	
Speaker10 (m)	457	1210		93	73	
Speaker11 (f)	221	540		98	72	
Speaker12 (m)	144	667		114	95	77
Speaker13 (f)	148	429		122	75	
Speaker14 (m)	333	646		97	93	60
Speaker15 (m)	248	548		80		
Speaker16 (f)	379	621		123	116	75
Speaker17 (m)		700		79	101	
Speaker18 (m)	367	826		123	89	
Speaker19 (f)	342	945		74	104	
Speaker20 (f)	239	484		103	76	81
Speaker21 (f)	288	699		112	145	153
Speaker22 (m)	345	1023		101	85	
Speaker23 (f)	325	782		109	87	
Speaker24 (m)	398	721		122	90	120
Mean of female dictors	285	699	1132	97	103	111
Mean of female dictors	318	725	967	130	128	159
Generalised Mean	**302**	**715**	**1024**	**115**	**118**	**135**

When analysing, with the help of Student t-test the data of foot lengthenings (cf. Table 2) we had to stick to the zero-hypothesis: the foot lengthenings are from samples of the same mean value.

Taken under scrutiny next, was whether and to what extent the prosodic parsing of speech correlates with syntactic parsing where the latter is indicated by punctuation marks and conjunctions. As evidenced in Table 3, in speech the pause[3] is invariably at every paragraph end and sentence end. Two third of commas are connected with pauses. The least marked in speech are phrases starting with such co-ordinating conjunctions, which do not require the comma.

[3] We have treated as a prosodic pause, in this work the interruption of speech over 50 ms.

Table 2. Student t-test results for comparing of two-sample means (Ph-Se – between phrase and sentence, Ph-Pa – between phrase and paragraph, Se-Pa – between sentence and paragraph)

	Pauses			*Foot lengthenings*		
	Ph-Se	*Ph-Pa*	*Se-Pa*	*Ph-Se*	*Ph-Pa*	*Se-Pa*
T stat	16,06	20,06	8,00	0,61	0,71	0,39
T critical two-tail	2,59	2,64	2,71	2,60	2,73	2,72
P(T<=t)	<0,0001	<0,0001	<0,0001	0,54	0,48	0,70

From among the punctuation marks, it is the dash that the lengthening is clearly connected with. Apparently, it is the spell of the form of the mark – the long line - that makes reader drawl. Suggestive of the link between pauses and boundary lengthenings is the English term „prepausal lengthening". The said term applies, on the basis of the Estonian language speech material, only in the extent of 60% (out of 601 pauses, the only 360 pauses displayed prepausal foot lengthening). According to perception tests carried out by Lehiste (Lehiste, Fox 1993) the Estonians expect a significantly lesser end lengthening on the last syllable of sentence that the English speakers do, as a matter of fact.

Table 3. Connection of pauses and foot lengthenings with the text parsing

	No of parsings in the text	*No of corresponding pauses in the speech*		*No of corresponding foot lengthenings in the speech*	
		Cnt	*%*	*cnt*	*%*
Paragrph end	21	21	100	15	71
Sentence end	185	185	100	124	67
Comma	179	120	67	94	53
Conjunction	85	42	49	46	55
Colon	11	10	91	6	57
Dash	15	13	87	14	93

4. Modelling of pauses and boundary lengthenings

Elucidation of mean durations of pauses and boundary lengthenings, in itself will not guarantee a natural rhythm of synthetic speech. On the basis of the analysis carried out in the previous section, in real speech the pauses and boundary lengthenings have a very large variance both in duration and also in their location in the speech flow. In order to preserve that variance in the synthetic speech, to some extent at least, the temporal structure of the speech must be modelled, according to context.

4.1. Modelling of durations

For modelling the pauses and boundary lengthenings, a vector of argument features (explanatory variables) consisting of 18 features was generated, basing on the text. Lumped under the vector were those features or factors that were likely to impact on duration of pause or foot lengthening.

The features give an indication of the structure of the text (e.g. sentence or phrase end, phrase length), prepausal or lengthened foot (e.g. foot quantity, foot length, length of last syllable of foot) and temporal structure of the speech (distance from the previous pause, distance from the intonation phrase). The response was the logarithmic

duration, which normalises the distribution of durations. The general parameters of the final model have been presented in Table 4. Quite clearly the model is significant and it described a large part of logarithmic duration variance (R-square=0,5564). Significant features impacting on pause duration are those related to punctuation marks and conjunction. A significant feature too is the distance from the previous pause and whether the foot preceding the pause has been lengthened. Construction of prediction model to pauses is seemingly successful, however modelling of foot lengthenings in the speech flopped, because the links with features were weak and the model described a small part only (R-square=0,1102) of variance of duration of lengthenings.

Table 4. Summary of fit and the analysis of variance for the regression model of durations

Summary of Fit					
Mean of Response	-0,86673		R-Square	0,5564	
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
Model	6	272,2	45,36	117,07	<.0001
Error	560	216,9	0,39		
C Total	566	489,1			

4.2. Modelling of situation of pauses in speech

When modelling the situation of pauses, we used the selection of argument features similar to that used in modelling the durations. Besides that we supposed that situation of the pauses may be in correlation with proper names and foreign words. We presumed that in front of the more complicated words in the speech flow, there could be a pause, and that after proper names we would tend to make interruptions in the speech (à la '*My name is Bond, James Bond*'). Anticipatively, that hypothesis failed to find proof on the basis of the given material. Response of the model was the value of probability that a pause would be made after a certain word.

*P(PAUS)= - 4.91 + 2,30 * FRKOM + 1,46 * FRSID + 3,80 * FRKLMK + 0,07 * KAUGLA + 0,21 * TAKTVXLDE + 0,20 * TAKPIKHX + 2,19 * PIKENDUS*

Figure 1. Model equation (P(PAUS) – probability value for a pause in the speech flow, FRKOM – phrase end (comma), FRSID – phrase end (conjunction), FRKLMK – phrase end (colon or dash), KAUGLA – distance from beginning of sentence in foots, TAKTVXLDE – quantity degree of the last foot, TAKPIKHX – length of last foot in sounds, PIKENDUS – feature of the lengthening of last foot)

The form of this binominal model has been presented in Figure 1. Here, too the situation of pause is in strong correlation with punctuation marks and co-ordinating conjunction. The probability of interruption of the speech flow is also heightened by distance from the beginning of sentence and whether the last foot was lengthened, as well as the last foot length and the quantity degree of the foot. The duration of the pause

will be calculated with the duration model found in p. 4.1. Admittedly prediction of lengthenings flopped, again. The model turned out inadequate.

5. Summary

Analysed in this work was the comportment of pauses and boundary lengthenings in the read out speech. In regression analysis, simple models for prediction of duration of pauses and their location in synthetic speech were found. For prediction of boundary lengthenings, no reliable prediction model could be derived. To all appearances, the lengthenings should not be treated in an isolated manner; rather they should be viewed as part of prediction model of sounds (Mihkla, Pajupuu, Kerge, Kuusik 2004). In further work, when composing the prediction models of pauses, different statistical methods too, should be used (e.g. neuron nets).

References

Eek, Arvo; Meister, Einar 2003. Foneetilisi katseid ja arutlusi kvantiteedi alalt (I): Häälikukestusi muutvad kontekstid ja välde. In: *Keel ja Kirjandu,* 11, 815–837.

Eek, Arvo; Meister, Einar 1998. Estonian Speech in the Babel Multilanguage Database: Phonetic-Phonological Problems Revealed in the Text Corpus. In: *Proceeedings of the Workshop on Speech Development for Central and Eastern European Languages.* The First International Conference on Language Resources and Evaluation, Granada.

Krull, Diana 1997. Prepausal lengthening in Estonian: Evidence from Conversational speech. In: Lehiste, I.; Ross, J. (eds.). *Estonian Prosody: Papers from a Symposium, Proceedings of the International Symposium on Estonian Prosody, Tallinn, Estonia, October 29-30, 1996.* Institute of the Estonian Language and Authors, Tallinn. 136–148.

Lehiste, Ilse 1981. Sentence and paragraph boundaries in Estonian. In: *Congressus Quintus Internationalis Fenno-Ugristarum, Turku, 20.-27. 1980, Pars VI.* 164-169.

Lehiste, Ilse; Fox, Robert 1993. Influence of duration and amplitude on the perception of prominence by Swedish listeners. In: *Speech Communication* 13, 149–154.

Mihkla, Meelis; Pajupuu, Hille; Kerge, Krista; Kuusik, Jüri 2004. Prosody modelling for Estonian text-to-speech synthesis. In: *The first baltic conference on human language technologies: the baltic perspective. Riga, April 21–22 2004,* Riga. 127–131.

Stout, Rex 2003. *Deemoni surm.* CD-versioon (loeb Andres Ots). Tallinn: Elmatar.

Zvonik, Elena; Cummins, Fred 2002. Pause duration and variability in read texts. In: *Proceedings of the 7th International Conference on Spoken Language Processing,* Denver, ICSLP-2002. 1109–1112.

MEELIS MIHKLA is assistant director, Institute of the Estonian Language, Tallinn. He received his M.A. (Estonian language) at University of Tartu, dealing with Estonian text-to-speech synthesis. His research interests concern prosody modelling and speech units databases. His doctoral study focuses on statistical modelling of temporal structure of speech.

IV

Mihkla, Meelis; Kerge, Krista; Pajupuu, Hille 2005.
Statistical modelling of intonation and breaks
for Estonian text-to-speech synthesizer. –
Proceedings of the 16th Conference of Electronic Speech Signal
Processing, joined with the 15th Czech-German Workshop
"Speech Processing", Robert Vich (Toim.), September 26–28.
Prague: 91–98, Dresden: TUDpress.

Statistical modelling of intonation and breaks for Estonian text-to-speech synthesizer[1]

Meelis Mihkla[1], Krista Kerge[2], Hille Pajupuu[1]

[1]Institute of the Estonian Language, [2]Tallinn University

Meelis.Mihkla@eki.ee

Abstract: For the synthetic speech to be acceptable to human ear, it is to feature the intonation, rhythm and accents sounding natural. The main objective of this work was to identify statistical models, basing on the read-aloud texts, which would help render natural the output speech in Estonian of the text-to-speech synthesizer. An attempt was made to find out, on the basis of one-to-one correspondence of the speech and the underlying text, whether and to what extent the prosodic parsing of speech is related to syntax and morphology. By regression analysis, values essential for predicting the fundamental frequency and sentence stress were elucidated. An effort to detect the links between speech parsing and breaks was undertaken.

1. Introduction

Text-to-speech system is expected to be able to generate the values of the fundamental frequency, which would not fall out with the characteristics of actual speech, to any noticeable degree. Modelling of the prosody of the current text-to-speech synthesis of Estonian [1] bases on a simple method of acoustic stylisation, where fundamental frequency in the sentence changes between the linear declination lines, wherefore the synthetic speech is much too monotonous. For that reason, the synthesizer does not enjoy a buoyant market demand, either. Speech contains complicated intonation, accent and rhythm patterns, which the text-to-speech system should be able to simulate, for the output speech to sound natural. In order to predict the complicated melody contours of the speech, various statistical methods are used. The intonation models built upon them have been drawn from databases of fluent speech, possessing the natural speech rhythm and rate [2, 3]. Such databases are not available for oral speech in Estonian, therefore up till now the researchers have been concerned solely with improvement of intonation of questions, on the basis of recordings designed for that particular end [4]. The goal of the present work is to find out whether the morphological and syntactic characteristics of text units, e.g. word class and form and its syntactic function, as well as position of the word in text (distance from the end of sentence and phrase[2]) are linked to prosodic characteristics of speech – sentence stresses[3], and values of fundamental frequency (peaks and valleys). An additional goal is to study where the breaks are positioned and when breathing occurs, as an organic part of naturalness of speech. Because we are concerned with text-to-speech synthesiser, serving as the source material are the read-aloud texts. On the basis of correspondence between written text and recorded speech, it is possible to move from

[1] This work was completed within the framework of Grant no. 5039 of Estonian Science Foundation, and the state programme "The Estonian language and the national memory".

[2] In this work, considered as phrase are the clause or element of enumeration, which have been delimited within the sentence by punctuation mark or conjunction.

[3] Marked here as words bearing the sentence stress are those, the peak of fundamental frequency whereof is higher than that of the word preceding it.

symbol presentation to acoustic presentation and to establish, whether and how much the morphological-syntactic parsing of text is related to prosodic parsing of speech. The material analysed derives from two fields: press (fragments of the news, volume 665 words) and fiction (passages from a detective story, 499 words). The press texts were read aloud by professional announcers (1 woman and 1 man), the fiction by a professional actor (man). The press text contained 44 sentences (average number of words in a sentence 15.1, SD 6.3), the fiction text 33 sentences (average 15.6, SD 8.9). The press sentences were found to be longer by ca. one word than a typical sentence; the fiction text, invariably more idiosyncratic, differed more from the typical one, i.e. by ca. 4 words; cf. Kerge [5]. All speech paragraphs were segmented into words and breaks. The morphological analysis was performed by means of a morphological analyzer [6]. The results were disambiguated and the syntactic functions found manually. (Automatic syntactic analyser is available for Estonian, however it was the original idea of the authors of this paper to find out, whether it is feasible to apply in speech synthesis several complicated and time consuming steps of automatic analysis.) Elected as primary method of statistical modelling of intonation, was the technology of regression models. The authors had previous experience in application of classical regression when predicting the temporal characteristics [7]. By use of statistical analysis, an attempt was made to optimise the number of characteristics, to be applied in prosody generator. The texts of different fields were analysed separately, because the fields are standing clearly apart, as regards the language units and their combination [8] and, judging by ear, also as regards the prosody and pattern of breaks.

2. Analysis and modelling the prosodic features

In the material subjected to morphological-syntactic analysis, breaks were subject to analysis in the first place, to be followed by study of correlation between form & function and prosody. Lastly, we modelled by regression analysis the values of peak and valley of the fundamental frequency in the word, also trying to predict the sentence stress.

2.1. Breaks and breathing

In recent years, ever more attention has been paid to breaks in running speech. Seemingly, breaks
are one of the most significant features in speech prosody [9]. Breaks are indispensable also to make the synthesised speech sound more natural. For that, one needs to know where and of what duration the breaks should be. According to Tseng [9] the breaks distribute hierarchically: the longest breaks mark the prosodic group boundary (in the text, it is reciprocated by the paragraph boundary). The prosodic group equals to at least one breath group, however it may contain multiple breath groups. The breath group includes, in its turn shorter breaks without breath. This work focused on situation and duration of breaks in speech in Estonian. After measurement of breaks, they were distributed into two subgroups: 1) breath group breaks; 2) breaks without breath (Fig. 1). With breath group breaks, both the duration of the break and the duration of breath were measured (Table 1)[4].

[4] Omitted have been the breath group breaks, coinciding with the prosodic group boundary.

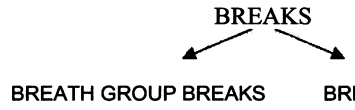


Figure 1: Types of breaks.

Given the fact that breath group breaks did not differ from one another in the dimension "female voice – male voice", but in the dimension "press text – fiction text", we concentrated on comparison of the press text with the fiction text. In the press text, the breath breaks are on average longer than those manifested when reading a fiction text. Although breathing takes more or less an equal amount of time, the duration of breath varies, when reading the fiction text, much more than in case of a press text.

Press text						**Fiction text**		
Male voice			Female voice			Male voice		
Break M/SD (ms)	**Breath** M/SD (ms)	Proportion of duration of breath in the duration of the whole break (%)	**Break** M/SD (ms)	**Breath** M/SD (ms)	Proportion of duration of breath in the duration of the whole break (%)	**Break** M/SD (ms)	**Breath** M/SD (ms)	Proportion of duration of breath in the duration of the whole break (%)
617/262	276/112	45%	597/256	269/99	45%	474/290	275/189	58%

Table 1: Average duration and range of breath group breaks.

In view of the text-to-speech synthesis we were keen to know to what extent the breaks (both breath group breaks and breaks without breath) coincide with punctuation marks of the text, cf. Table 2.

Press text				**Fiction text**		
	Number in text	**Number of breaks**	**Breath**	**Number in text**	**Number of breaks**	**Breath**
Full stop	44	44	36	33	32	18
Comma	**37**	**27**	12	**41**	**16**	13
Dash	8	8	6	5	4	1
Colon	6	6	5	1	1	0
Semicolon	1	1	1	0	0	0
Break without a punctuation mark		**19**	6		**38**	15

Table 2: Correspondence of breaks with the punctuation marks.

Again, the difference in reading of fiction text and press text should be pointed out: in the press text, one generally tends to abide by the punctuation marks. The sole compromise allowed may be failing to observe the break in the place of comma (27% of cases). When reading the press text, as compared with the fiction text, the additional breaks occur less frequently by half in places, featuring no punctuation marks. Breathing, too predominantly follows the punctuation marks, when reading the press text, while additional breaks are used less frequently.

When reading the fiction text, the punctuation marks are less observed, for making breaks. In the place of comma, the break is not observed in 61% of cases. To compensate, the speech features a lot of additional breaks. Breathing occurs rather in the place of commas than full stops, and also during additional breaks. In press texts the places where additional breaks are

created are relatively well regulated by rules: the reader tends to make additional breaks in front of conjunction *ja/ning* ('and'), in front of names and figures. In fiction, the places where additional breaks are created are mainly meaningful, i.e. not subject to rules, without comprehension of the content of the text. Cf. Table 3.

Additional break	Press text	Fiction text
Conjunction in front of *ja/ning* ('and')	5	2
In front of name	2	2
In front of figure	5	1
Total additional breaks in texts	19	38

Table 3: Places where the additional breaks are created.

When reading the press and fiction texts, the main difference concerning the breaks seems to be, briefly that the readers of the press text observe, while making the breaks, rather dutifully the form of the text (punctuation, numerals written in figures, names with capital letters), the readers of fiction texts proceeding rather from the content of the text than from its form. Hence the breaks of the synthesised speech are easier to regulate by rules than basing on data of press text.

Taking into consideration the parsing of press text, punctuation, breathing of the reader and additional breaks we obtain 6 break groups, varying in duration: 1) prosodic group break (paragraph boundary) with breath; 2) full stop break with breath; 3) colon, dash, semicolon break with breath; 4) comma break with breath; 5) comma break without breath; 6) break without punctuation mark without breath (in front of *ja* 'and', name, figure). The durations of the break groups are presented in Table 4.

Break groups	Average duration of the break (ms)	Average duration of the breath (ms)
Prosodic group break	988	330
Full stop break	613	273
Colon, dash, semicolon break	486	227
Comma break	398	222
Comma break	170	0
Break without punctuation mark (in front of *ja* 'and', name, figure)	69	0

Table 4: The average duration of the break groups.

Since breathing entails elevation of fundamental frequency (start of a new breathing group), the expected duration of breath groups should be taken into consideration when electing comma break with or without breath. The average duration of 96 breath groups under scrutiny was 4278 ms, SD 1985 ms.

The outcome of research of breaks corroborates the impression obtained by ear that the fields (press and fiction) are to be analysed separately.

To all evidences, the press text may be regarded as standard of the readable text, a basis of com-parison with other texts – the other data available, too suggest that the press represents the average indicators of the textual area of the Estonian language [5] and is therefore accepted asneutral. The neutral reading befits all genres (also the reading of fiction), while the contrary is not the case.

2.2. Modelling of intonation

The main objective of modelling the intonation was verification by statistical methods, whether and to what extent one should apply the morphological and syntactic analyser in speech synthesis, when generating the values of fundamental frequency of the word, and the sentence stresses. For that we tried to model the peak values of fundamental frequency of the word and the value of word final fundamental frequency, as well as sentence stresses, on the basis of morphological - syntactic information of the text. Basing on text, for every word, the vector of features was generated, containing 15 coded features characterising the word class, 11 – the word form, 13 – the syntax, 2 – the position of the word[5], 1 – the duration of the word, 2 – the fundamental frequency contour of the word. Output of the model or response – fundamental frequency – was presented on the scale of semitones. Because the argument features were numerous (total 44 features), an optimum choice was to be made from among themselves, i.e. one had to identify the features having the greatest impact on output. Firstly, an estimate was given to the link of argument features to response, on the basis of correlation matrix. From there, 13–15 features were selected for the models of multiple regression analysis. The values of peaks and valleys of fundamental frequency were modelled separately. Presented as an example in Fig. 2 have been the male announcer regression equations for calculation of fundamental frequency peak and valley values in the word.

$$Peak = 1.90*SAL + 3.16*SAF + 4.58*RQHK - 0.83*T$$

$$Valley = 2.79*SAF - 0.002*KEST - 1.20*ID - 1.85*inf + 2.60*J$$

Figure 2: Fundamental frequency peaks and valleys regression equations for male announcer (*SAL – position of word in sentence, SAF – position of word in phrase, RQHK – stressed nature of word, T – attribute, KEST – duration of word, ID – uninflected word form, inf –infinitive with da-feature, J – conjunction*).

Presented as an example in Fig. 3 have been, the position of actual fundamental frequency contour peaks and reckoned peaks in three sentences. In order to predict the position of sentence stress, the binominal regression model was composed. Output of the model is the probability value of stressed nature for every word (if P(stress) < 0.5, this word is not stressed in the sentence, if P(stress) ≥ 0.5, the sentence stress is on that word). Presented in Table 5 have been important features in prediction of fundamental frequency peak and valley and determination of sentence stress position as per announcers (sign + or – in front of the feature indicates in which direction the said feature impacts on the output). For instance, word final fundamental frequencies of the actor are predominantly determined by the position of word in sentence, while frequencies of word final fundamental frequency of pre- or postpositions fall lower than those of other words. All models corresponding to features presented in Table 5 are statistically meaningful. The most pliant to modelling were the values of fundamental frequency of peaks, and the weakest links were manifested in determining the position of sentence stresses. The prediction models of fundamental frequency of peaks also described a sizeable share of variance of peak frequencies of fundamental frequency (r^2=0.565), cf. Table 6.

[5] Position of the word in the sentence or phrase was determined by the so-called normalised distance from the sentence or phrase end, i.e. at the beginning of sentence or phrase the value of distance is 1.0 and at sentence or phrase end distance is 0.0

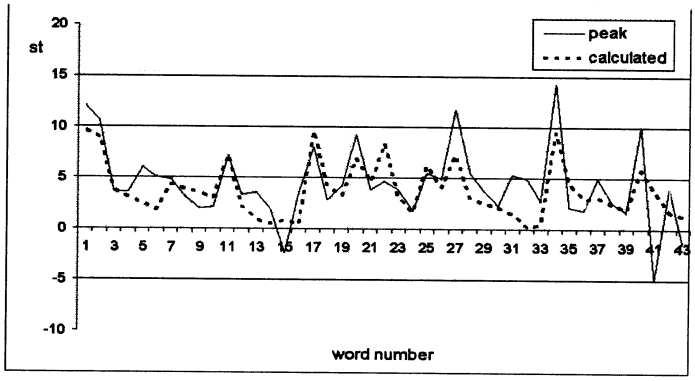


Figure 3: Fundamental frequency peaks contour and calculated contour of the news text fragment read by male announcer.

Speakers	Peaks	Valleys	Sentence stress
Actor (male)	+ Position of word in sentence – Position of word in phrase + Sentence stress – Pronoun; – Predicate	+ Position of word in sentence – Pre- or postposition	+ Word duration
Announcer (female)	+ Position of word in sentence + Position of word in phrase + Sentence stress, – Adverb	+ Position of word in sentence + Position of word in phrase – Word duration	+ Substantive + Adjective + Subject
Announcer (male)	+ Position of word in sentence + Position of word in phrase + Sentence stress – Attribute	+ Position of word in phrase – Uninflected word form – Infinitive with *da*-feature + Conjunction + Adverb; – Word duration	+ Duration – Reverse intonation

Table 5: Important features in prediction of fundamental frequency and sentence stress as per announcers (sign '+' in front of the feature means that the said feature increases the value of the output, while the sign '–' means that it decreases the same.

Summary of Fit					
Multiple R 0.758			R-Square 0.5393; Adj R-Sq 0.5368		
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
Model	14	2105.5	150.4	26.87	<0.0001
Error	301	1681.8	5.08		
C Total	315	3086.2			

Table 6: Summarised suitability of regression model of fundamental frequency of peaks and the variance analysis, in case of male announcer.

In modelling the fundamental frequency, the regression equations clearly reveal the declination lines: both on the sentence level (SAL – position of word in sentence) and within the phrase (SAF – position of word in phrase). The phrase parameter is absent only in the valleys equation of the actor. The parameters of declination lines are adjusted for a particular word by word class and word form features and by function of word duration, stressed nature and syntactic function.

For instance, the peaks of pronoun and attribute are lower than with other word classes. The stressed nature of the word increases significantly the value of peak, while the fundamental frequency of word final will remain the same. In the pre- or postposition, the uninflected word form and the infinitive with *da*-feature, the valley is lower than the declination line. In longer words the fundamental frequency falls lower in valleys than in words shorter by duration, while with conjunctions the fundamental frequency drops less, during the progress of the word. Unfortunately, the majority of those links with intonation rather manifest the individual characteristics of the announcer's speech, not enabling identification of more general links.
As was to be expected, stressed nature of word in the sentence is the hardest to predict. As it is, everybody can freely interpret the text, hence the impressive variance of position of sentence stresses. The speech of the female announcer is characterised by consistent tendency to stress substantives and adjectives and the words, syntactically functioning as the subject. In the speech of male announcers, the stressed words are longer, by duration.

3. Summary and discussion

Breaks occupy an important place in providing natural rhythm to synthesised speech. In the material researched, breaks are inserted for purely syntactic reasons in press text, only – the fiction material suggests that due to its breaks prompted by meaning it is not suitable as a model; it also suggests that for improving the sound of synthesised speech, too the press texts might be of help. Inserting breath breaks in output speech, after the pattern of press text, helps to make the synthesised speech livelier. Modelling of intonation revealed clearly the sentence and phrase level declination lines. The morphological-syntactic information specified the values of peaks and valleys in case of a particular word, however the more general link between fundamental frequency and syntax was lacking. Prediction of stress was affected by the fact that all announcers had different texts and therefore, detection of regularities may have been impeded. Thus, almost no correlations between prosody and morphological form or between prosody and syntax turned out statistically meaningful, in particular those, which would manifest themselves in case of all presenters of text or both fields (cf. Table 5). Therefore, in the first approximation one might assume that it does not make sense to apply the syntactic analyser in full, for synthesis in Estonian. This work however was but one of the first tentative steps in finding links between the prosodic features and syntactic parsing of the text. That outcome was both anticipated and affronting. On the one hand, the negative outcome is due to the multiplicity of morphological forms of Estonian and the resulting free word order. Although there are three main sentence types in Estonian, the doctoral paper by Kaja Tael of not too recent past [10] points out in the press text sentence ca. n word order patterns. On the other hand, intonation, stress and insertion of breaks must be related to syntax, too in case of free word order. Hence some links found and not taken into account in current modelling, as statistically little meaningful, suggest the sensibleness of composition and analysis of a larger database.
For instance sentence stress has been, in all cases studied, i.e. both in case of fiction text and press text, independent of reader related to two features studied, with substantives as word class (whose functions however have been partly underrepresented) and with subject of the sentence. Whereas the link between peak and pronoun is negative, which suggests the need to more accurately take into account the word classes and their sentence function – e.g. the personal pronoun is relatively frequently in the function of subject.
Besides the stressed nature of substantives, in case of two male voices (independent of the field of text), fundamental frequency elevates over them. For readers of press text, the noun is, in its turn in negative correlation with the valley of fundamental frequency. Statistically relatively important is also the fact that the noun, as a stressed component in the material studied is preceded by something, i.e. it has mostly attributes. Regarding the attributes, only

in case of a female voice a negative link with the peak is manifested (as statistically meaningful, at that), wherefore it would be fascinating to study whether the outcome would be the same if the female voice were to read the prose of a different field, e.g. fiction, and what would be the result of a male voice with the same material. Male voices are also united by the fact that both the conjunction as word class and the connector as member of sentence are in the valley of intonation. With the press text, both with male and female voices it is the complement that occurs in valley. It is not trivial, either that in all cases studied the stress is in negative correlation with uninflected form of the word, independent of its class, whereas the uninflected form is, in case of two voices (in case of a male presenting fiction text and a female presenting press text), in correlation with the valley. With the actor, the predicate is negatively related to the peak of intonation, in case of both announcers, however the signs of analogical trend are obvious: with the female voice, verbs as the word class fall in the valley, and with male voice – part of the declining forms of the verb fall there. In the further work it would be reasonable to extend the volume of data analysed, to specify the choice of features and to attempt to apply in modelling different statistical methods (generalised regression, decision trees, neural networks.

References

[1] Mihkla, M.; Meister, E.: Eesti keele tekst-kõne süntees. Keel ja Kirjandus, 2 (2002), 88– 97, 3 (2002), 173–182.

[2] Campbell, N.: Timing in speech: a multilevel process. In: Horne, M. (ed.) Prosody: theory and experiment. Dordrecht/Boston/London: Kluwer Academic Publishers, 2000, 281– 334.

[3] Vainio, M.: Artificial neural network based prosody models for Finnish text-to-speech synthesis. 2001. University of Helsinki, Helsinki.

[4] Mihkla, M.; Pajupuu, H.; Kerge, K.: Modelling and perception of the Estonian general questions with the *kas*-particle. Proceedings of 15th ICPhS. Barcelona, 2003, 539–542.

[5] Kerge, K.: Kirjakeele kasutusvaldkondade süntaktiline keerukus. In: Kasik, R. (ed.) Tekstid ja taustad. Artikleid tekstianalüüsist. Tartu Ülikooli eesti keele õppetooli toimetised 23. Tartu Ülikooli Kirjastus, 2002, 105–119.

[6] EKI morfoanalüüs ja süntees 3.2. http://www.eki.ee/keeletehnoloogia/projektid/morfana/

[7] Mihkla, M.; Pajupuu, H.; Kerge, K.; Kuusik, J.: Prosody modelling for Estonian text-to-speech synthesis. The Proceedings of the First Baltic Conference. Riga, 2004, 127–131.

[8] Meier, H.: Essee asend tekstitüübivõrdluses. (Magistritöö. Käsikiri Tallinna Ülikooli raamatukogus.) Tallinn: TPÜ, 2003.

[9] Tseng, C.: The prosodic status of breaks in running speech: examination and evaluation. Speech Prosody 2002, Aix-en-Provence, France, 2002, 667–670.

[10] Tael, K.: Sõnajärjemallid eesti keeles. Doctoral thesis. Tallinn, 1988.

V

Mihkla, Meelis 2006.
Pausid kõnes. Keel ja Kirjandus, XLIX(4): 286–295.

PAUSID KÕNES

MEELIS MIHKLA

Kõnes me pausidele teadlikult olulist tähelepanu ei pööra. Kõneprosoodias on aga pausidel oluline roll. Pausid korraldavad kõne esmast süntaktilist liigendust või prosoodilist fraseerimist eesmärgiga kergendada lausungist arusaamist ja nad on üks kõnerütmi kujundavaid tegureid. Praeguse eestikeelse tekst–kõne-sünteesi reeglipõhises prosoodiageneraatoris[1] pause ei modelleerita, vaid nad on jäikade reeglitega fikseeritud, mis võib olla üks põhjusi, et väljundkõne on monotoonne ja ebaloomuliku rütmiga. Käesoleva töö eesmärgiks on etteloetud erinevat tüüpi tekstide salvestiste põhjal analüüsida pauside kestusi ning nende asukohti sidusas kõnes ja modelleerida neid eestikeelsele tekst–kõne-sünteesile.

Et tehiskõne tunduks inimkõrvale loomulik, peaks ta sisaldama loomuliku kõlaga intonatsiooni, rütmi ja rõhuasetust. Ehk täpsemalt, tekst–kõne-süsteem peab olema võimeline genereerima selliseid häälikute ja pauside kestusi ning põhitooni väärtusi, mis ei erine oluliselt reaalse kõne vastavatest väärtustest. Foneetikas ja fonoloogias on pausidele seni suhteliselt vähe tähelepanu osutatud. Suulise kõne lingvistilistes uurimustes on kõneüksustena käsitletud häälikuid, silpe, kõnetakte, sõnu ja fraase põhiliselt isoleeritud lause koosseisus. Lausesiseselt on aga pause raske käsitleda toimivate kõneüksustena, mis võibki olla peapõhjuseks, miks neid on seni lingvistilis-foneetiliselt tähtsusetuteks peetud.[2] Viimasel kümnendil, kui foneetilises uurimistöös on hakatud laialdaselt kasutama kõnekorpusi, on pausidele kui kõneprosoodia olulisele tunnusele järjest enam tähelepanu pööratud.

Pausid kõnes

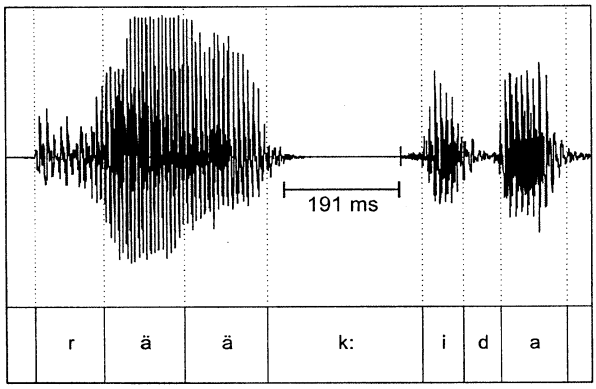
Kõnes esinevate pauside klassifitseerimisel on kaks põhilist lähenemisviisi: füüsikalis-lingvistiline ja psühholingvistiline jaotus.[3] Traditsioonilises lingvistilises pausi määratluses peetakse kõnevoogu füüsikalise pausiga katkestatuks, kui on täheldatav hääletus ehk vaikus akustilises signaalis, s.t selles kõnesegmendis on kõnesignaal ilma märkimisväärse amplituudita. Lingvistilisest kontekstist sõltuvalt on kõnes kaht liiki hääletuid pause:

– häälikusisesed pausid, näiteks eesti keele klusiilides esineb paus kõnetrakti sulufaasis (vt joonist 1);
– sõnadevahelised pausid.

[1] M. Mihkla, E. Meister, A. Eek, Eesti keele tekst-kõne süntees: grafeem-foneem teisendus ja prosoodia modelleerimine. Arvutuslingvistikalt inimesele. TÜ üldkeeleteaduse õppetooli toimetised 1. Tartu, 2000, lk 309–320.

[2] C. Tseng, The Prosodic Status of Breaks in Running Speech. Examination and Evaluation. – Proceedings of Speech Prosody 2002. Aix-en-Provence, 2002, lk 667–670.

[3] B. Zellner, Pauses and the Temporal Structure of Speech. – Fundamentals of Speech Synthesis and Speech Recognition. Ed. E. Keller. Chichester: John Wiley, 1994, lk 41–62.



Joonis 1. Sõna *rääkida* kõnelaine, kus on näha häälikusisene paus 191 millisekundit.

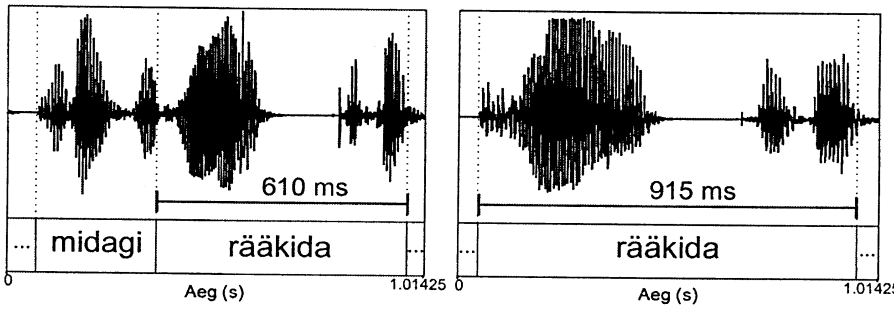
Kui häälikusisene paus on klusiilide oluline segment, siis sõnadevahelised pausid korraldavad kõne esmast liigendamist või prosoodilist fraseerimist eesmärgiga kergendada lausungi tajulist tõlgendamist. Sõnade vahel pausi ajal toimub ka enamik sisse- ja väljahingamisi. On oluline märkida, et see, mida meie kõrv tajub pausina, ei ole vaid kõnes tuvastatav vaikus. Tegelikult on pausi tajumine signaali tasandil kestuse, põhitooni ja kõnesignaali intensiivsuse keerukas kombinatsioon. Psühholingvistikas käsitletakse pausi kui suprasegmentaalset nähtust, sest pausiga on seotud palju kõnesegmente ja paus sõltub suure hulga füüsikaliste parameetrite koosmõjust. Selle lähenemisviisi põhjal jagatakse pausid

– hääletuteks pausideks, mis vastavad kõnesignaalis tajutavale vaikusele;
– "täidetud" pausideks ehk kõnetakti pikendusteks.

Enamik "täidetud" pause on seotud lause- ja fraasilõpu viimaste silpide ja kõnetaktide venituse-pikendusega. Selliseid n-ö auditiivse illusiooni pause nimetatakse lõpupikendusteks (vt joonist 2). Lõpupikendused ehk pausieelsed pikendused kannavad endas infot, et pikendatud kõnetaktiga sõnale võib järgneda tegelik paus. Peale lõpupikenduste esineb kõnes selliseid fraasisiseseid kõnetakti pikendusi, mis on seotud esiletõsterõhuga või fokuseerimisega, s.t fraasi teatavaid sõnu tõstetakse kõnes esile nende kestust pikendades.

Ehkki pauside kõnevoos paiknemine ja nende klassifitseerimine näitavad, et eri keeltes võib olla ühiseid seaduspärasusi, on uurimused välja toonud ka märkimisväärseid erinevusi, näiteks itaaliakeelses kõnes on pausid oluliselt lühemad kui hispaania keeles.[4] Näide 1 illustreerib pauside paiknemist eestikeelses kõnevoos. Võrdlusena on vasakus veerus etteloetud tekst ja paremal vastava kõnevoo lihtsustatud esitlus: pausid grafeemijadas. Näeme, et teksti struktuur on oluliselt rangem: üldjuhul on iga sõna lõpus tühik ja iga lause lõpus kirjavahemärk. Kõnes võib iga inimene teksti küllalt vabalt interpreteerida: sõnadevahelised pausid paiknevad sõnarühma või prosoodilise fraasi järel, aga prosoodilised fraasid ei pruugi kokku langeda süntaktiliste fraasidega ja lõpupikendustel on tendents paikneda prosoodilise fraasi lõpus, kuigi mitte alati. Osa näites 1 allajoonitud pikendatud kõnetakte on seotud

[4] E. Campione, J. Véronis, A Large-Scale Multilingual Study of Silent Pause Duration. – Proceedings of Speech Prosody 2002, lk 199–202.



Joonis 2. Ühesuguse pikkusega lõigud kõnelainest, kus sõna *rääkida* on erinevas kontekstis erineva pikkusega. Vasakpoolsel kõnelainel on *rääkida* lausesiseses positsioonis (kestus 610 millisekundit), parempoolsel kõnelainel paikneb sama sõna lause lõpus kestusega 915 ms, mis viitab lõpupikendusele ehk pausieelsele pikendusele ja mis kannab infot, et sellele võib järgneda paus.

Talle meeldis nendega uhkustada – kui need teie omad oleksid, meeldiks see teilegi –, aga mitte sellepärast ei seganud ta vahele. Ta tahtis paari kirja dikteerida ja ta arvas, et kui ma missis Hazeni üles orhideesid vaatama viin, siis ei tea keegi, millal me sealt alla tuleme. Aastaid tagasi jõudis ta ebapiisavatele tõenditele tuginedes otsusele, et ma kaotan veetlevate noorte naiste seltskonnas ajataju, ja kui tema kord midagi otsustab, siis on see otsustatud.	Tallemeeldisnendegauhkustada**P** kuineedteieomadoleksidmeeldiksseetei legi**P**agamittesellepärasteiseganudtava hele**P**Tatahtis**P**paarikirjadikteeridaja taarvasetkuimamissisHazeniülesorhi deesidvaatamaviinsiiseiteakeegi**P**millal **P**mesealtallatuleme**P**Aastaidtagasijõu dista**P**ebapiisavatele tõenditeletuginede sotsuseleetmakaotanveetlevatenoorte naiseltskonnasajataju**P**jakuitema kordmidagiotsustabsiisonseeotsustatud **P**

Näide 1. Etteloetud teksti struktuur *versus* pausid kõnevoos. Vasakul veerus etteloetud tekst ja paremal veerus kõnevoos olevad pausid (**P** – sõnadevahelised pausid).

fokuseerimisega (näiteks fraasis *veetlevate noorte naiste seltskonnas* on esile tõstetud sõna *naiste* kõnetakti pikendusega).

Järgnevalt keskendutakse sõnadevahelistele hääletutele pausidele (edaspidi tähistame pausina just seda pausi tüüpi) ja lõpupikendustele ehk pausieelsetele pikendustele.

Eestikeelse kõne pause ja lõpupikendusi on uuritud põgusalt või riivamisi teiste ülesannete kontekstis. Ilse Lehiste on kontrollinud, kas lõpupikendused on korrelatsioonis järgnevate pauside pikkustega, ja tuvastanud väga nõrga seose.[5] Diana Krull on uurinud pausieelseid pikendusi dialoogkõnes kahesilbilistes sõnades väldete kontekstis.[6] Arvo Eek ja Einar Meister on mõõt-

[5] I. Lehiste, Sentence and Paragraph Boundaries in Estonian. – Congressus Quintus Internationalis Fenno-Ugristarum, Turku, 20.–27. VIII 1980, Pars VI. Turku, 1981, lk 164–169.

[6] D. Krull, Prepausal Lengthening in Estonian. Evidence from Conversational Speech. – I. Lehiste, J. Ross (eds.), Estonian Prosody: Papers from a Symposium. Proceedings of the International Symposium on Estonian Prosody. Tallinn, Estonia, October 29–30, 1996. Tallinn: Institute of the Estonian Language, 1997, lk 136–148.

nud lauselõpu pikendusi tempokorpuse baasil.[7] Aga ka neil olid vaatluse all vaid kindla struktuuriga sõnad ja põhitähelepanu keskendus väldete tunnustele. Eestikeelse tekst–kõne-sünteesi jaoks on vaja mõõta sidusa kõne pause ja lõpupikendusi.

Algmaterjal

Et uurimuse eesmärgiks on analüüsida ja modelleerida pauside kestusi ning nende asukohti sidusas kõnes eestikeelsele tekst–kõne-sünteesile, siis on valitud lähtematerjaliks diktorite etteloetud erinevat tüüpi tekstid. Teksti ja kõne üksühese vastavuse põhjal saab prosoodia sümbolesituselt üle minna akustilisele ning samuti tuvastada, kas ja kuivõrd on teksti süntaktiline liigendus seotud kõne prosoodilise liigendusega.

Lähtematerjaliks võeti kõnelõigud näitleja etteloetud kriminaalloo CD-versioonist,[8] kõnelõigud ning tekstid Eesti Raadio pikematest diktorite loetud uudistest ja kõnelõigud eesti foneetilisest andmebaasist BABEL.[9]

Kokku on analüüsitud 44 kõnelõiku (igaüks 0,5–2 minutit pikk) 27 diktori (14 mehe ja 13 naise) esituses. Diktorid lugesid erinevaid kõnelõike, vaid BABEL-i andmebaasi salvestiste korral lugesid ühte ja sama teksti 2–3 diktorit. Kõik kõnelõigud on segmenteeritud häälikuteks ja pausideks.

Eestikeelse kõne pausid ja pausieelsed pikendused

Töös analüüsitakse eelkõige neid pause ja lõpupikendusi, mis on seotud kirjavahemärkide ja sidesõnadega. Selleks mõõdeti etteloetud tekstide kõnelainetest pauside kestused ja arvutati kõnetakti pikendused. Kõnetakti pikenduste arvutamiseks summeeriti kõnetakti moodustavate häälikute kestused ja võrreldi saadud summeeritud kestust sama taktistruktuuri keskmise kestusega konkreetse diktori kõnes. Peale struktuuri arvestati ka taktiväldet. Kui mingi taktistruktuur osutus antud tekstis unikaalseks struktuuriks (nt CVCCC-CV sõna *korstna*), siis võrreldi selle kestust mingi sarnase kõnetakti struktuuriga (nt CVCC-CV sõnaga *kordse*, lahutades *korstna* hääliku kestuste summast konsonantühendi ühe komponendi kestuse).

Järgnevalt vaatleme, kas ja kuivõrd erinevad omavahel kestuse poolest fraasi-,[10] lause- ja lõigulõpu pausid ja vastavad taktipikendused. Tabelis 1 on pauside[11] ja lõpupikenduste keskmised kestused diktorite kaupa, mees- ja naisdiktorite keskmised ning üldine keskmine. Lõigulõpu pauside väärtused puuduvad BABEL-i andmebaasi diktorite kõne puhul, sest seal etteloetav tekst piirdus ühe lõiguga. Tabelist on näha, et isegi keskmiste väärtuste variatiivsus on väga suur. Huvitav on siiski märkida, et meeste ja naiste pauside üldised keskmised erinevad kestustelt üksteisest vaid 10% piires.

[7] A. Eek, E. Meister, Foneetilisi katseid ja arutlusi kvantiteedi alalt (I): Häälikukestusi muutvad kontekstid ja välde. – Keel ja Kirjandus 2003, nr 11, lk 815–837.

[8] R. Stout, Deemoni surm. CD-versioon (loeb Andres Ots). Tallinn: Elmatar, 2003.

[9] A. Eek, E. Meister, Estonian Speech in the BABEL Multi-Language Database. Phonetic-Phonological Problems Revealed in the Text Corpus. – Proceedings of LP'98. Vol II. Ed. O. Fujimura *et al.* Prague: The Karolinum Press, 1999, lk 529–546.

[10] Selles töös käsitletakse fraasina osalauset või loetelu elementi, mis on lausesiseselt piiritletud kirjavahemärgi või sidesõnaga.

[11] Prosoodilise pausina ja lõpupikendusena on selles töös käsitletud kõne katkestust ja lõpupikendusele vastavat kõnetakti pikendust, mille kestus ületab 50 millisekundit.

Tabel 1.

Pauside ja lõpupikenduste keskmised kestused (millisekundites) diktorite kõnes

Diktorid	*Fraasilõpu pausid*	*Lauselõpu pausid*	*Lõigulõpu pausid*	*Fraasilõpu pikendused*	*Lauselõpu pikendused*	*Lõigulõpu pikendused*
Näitleja (mees)	352	558	1025	200	220	315
Raadiodiktor (m)	303	828	902	124	112	117
Raadiodiktor (n)	286	769	1132	95	90	122
Diktor1 (naine)		547		103	107	113
Diktor2 (m)	361	862		60	73	77
Diktor3 (n)	255	306		76	138	
Diktor4 (m)	145	478		78		89
Diktor5 (n)	275	879		109	100	88
Diktor6 (n)	470	1179		89	89	
Diktor7 (n)	117	559		85		64
Diktor8 (m)	416	829		73	118	75
Diktor9 (m)	260	683		94	91	
Diktor10 (m)	457	1210		93	73	
Diktor11 (n)	221	540		98	72	
Diktor12 (m)	144	667		114	95	77
Diktor13 (n)	148	429		122	75	
Diktor14 (m)	333	646		97	93	60
Diktor15 (m)	248	548		80		
Diktor16 (n)	379	621		123	116	75
Diktor17 (m)		700		79	101	
Diktor18 (m)	367	826		123	89	
Diktor19 (n)	342	945		74	104	
Diktor20 (n)	239	484		103	76	81
Diktor21 (n)	288	699		112	145	153
Diktor22 (m)	345	1023		101	85	
Diktor23 (n)	325	782		109	87	
Diktor24 (m)	398	721		122	90	120
Naisdiktorite keskmine	285	699	1132	97	103	111
Meesdiktorite keskmine	318	725	967	130	128	159
Üldine keskmine	**302**	**715**	**1024**	**115**	**118**	**135**

Üldiste keskmiste visuaalse vaatluse põhjal võib arvata, et normaalse kõnetempoga etteloetud teksti puhul on pausid kestuse poolest eristatavad. Seda kinnitab ka valimite statistiline analüüs. Analüüsides paarikaupa pauside logaritmitud kestuste keskmisi Studenti t-testiga, tulid t-statistiku väärtused {16,06; 20,06; 8,00} olulisuse nivool $p = 0,01$ märgatavalt suuremad t-jaotuse kahepoolsest kvantiilist {2,59; 2,64; 2,71} hüpoteesi olulisuse tõenäosusel $P < 0,00001$. Seega võib pidada tõestatuks, et pauside kestuste keskväärtused erinevad ning et pauside klassifikatsioon on nende kestuste alusel võimalik, ja seda võiks kõnesünteesis rakendada. Väärtuste variatiivsus oli aga niivõrd suur, et näiteks kõnetuvastuses ei ole sellise klassifikatsiooniga suurt midagi peale hakata.

Analüüsides Studenti t-testiga taktipikenduste andmeid, tuli jääda nullhüpoteesi juurde: kõnetakti pikendused olid ühesuguse keskväärtusega valimitest.

Teise sammuna oli vaatluse all, kas ja kuivõrd on kõne prosoodiline liigendus ja teksti süntaktiline liigendus korrelatsioonis seal, kus süntaktilist liigendust tähistavad kirjavahemärgid ja sidesõnad. Nagu tabelist 2 näha, on kõnes paus alati iga lõigu lõpus ja peaaegu iga lause lõpus. Vaid näitleja lubas endale vabaduse kõnes kaks lauset kokku lugeda. Väga tugev on süntaksi ja prosoodia seos ka kooloni ja mõttekriipsu korral. Kaks kolmandikku komadest on seotud pausidega. Kõige vähem markeeritakse kõnes nende rinnastavate sidesõnadega algavaid fraase, mis üldjuhul koma ei nõua (*ja, ning, ega, ehk, või, kui ka*).

Tabel 2.

Pauside ja lõpupikenduste ning teksti liigenduse seos

	Liigenduste arv tekstis	*Liigendusele vastavate pauside arv kõnes*		*Liigendusele vastavate lõpupikenduste arv kõnes*	
		Arv	*%*		*%*
Lõigu lõpp	21	21	100	15	71
Lause lõpp	185	184	99	124	67
Koma	179	120	67	94	53
Sidesõna	85	42	49	46	55
Koolon	11	10	91	6	57
Mõttekriips	15	13	87	14	93

Lõpupikendusega on kirjavahemärkidest selgeim seos mõttekriipsul. Ilmselt tingib selle lugeja jaoks juba märgi kuju ise: pikk kriips kutsub esile sõnade venitamise. Pauside ja lõpupikenduste omavahelisele seotusele viitab inglise keelest pärit termin *pausieelne pikendus* (*prepausal lengthening*). See termin kehtib analüüsitud eestikeelse kõnematerjali põhjal vaid 60% ulatuses (601 pausist oli eelneva taktipikendusega vaid 360 pausi). I. Lehiste läbiviidud tajutestide põhjal[12] eeldavadki eestlased lause viimasel silbil oluliselt väiksemat lõpupikendust kui näiteks inglise keele kõnelejad.

Pauside modelleerimine

Eelnev analüüs näitas, et pausidel on kõnes väga suur variatiivsus, kuid eri liiki pausid on kestuse poolest eristatavad. Vaevalt et sünteeskõne rütm ja loomulikkus sellest oluliselt paraneks, kui me iga teise koma järel ja iga kolmanda sidesõna ees teeksime konstantse, fraasilõpu pausi. Kõne loomulikkus pigem eeldaks, et me oskaksime nii pauside kestuse kui ka kõnevoos paiknemise variatiivsust sünteeskõnes mõistlikult edasi anda. See ei ole lihtne ülesanne. Kõnepauside suurt variatiivsust on püütud vähendada ja modelleerimisülesannet lihtsustada sellega, et algmaterjalina on kasutatud sünkroonlugemist, mil kaks või enam diktorit loevad üheaegselt ette ühte ja

[12] I. Lehiste, R. Fox, Influence of Duration and Amplitude on the Perception of Prominence by Swedish Listeners. – Speech Communication 1993, nr 13, lk 149–154.

sedasama teksti.[13] Sellisel lugemisel teevad diktorid kõnevoos pause enamvähem samal ajal ja sarnase kestusega. Paraku on sünkroonkõne kõlalt ebaloomulik ja tal on tendents muutuda skandeerimiseks.

Selles töös on modelleerimise algmaterjalina kasutatud põhiliselt samu kõnelõike mis analüüsilgi. Mudelite treenimisvalimist jäeti välja ilukirjandusele vastavad kõnelõigud, sest ilukirjandusliku teksti fraseerimine erineb näitlejate kõnes tunduvalt näiteks uudiste lugemisest ning nõuaks erikäsitlust. Ka on sünteesi väljundkõne praeguse kvaliteedi juures süntesaatorilt vara nõuda ilukirjanduse ilmekat ettelugemist.

Pauside kestuste modelleerimiseks genereeriti teksti põhjal tunnused, mis kirjeldasid:

– teksti struktuuri (lõigu-, lause- ja fraasilõpp, sidesõnad);
– pausile eelnevat kõnetakti (takti pikkus häälikutes, taktivälde, takti viimase silbi pikkus häälikutes ja binaarne tunnus, mis näitas lõpupikendust);
– pausi ajalisi suhteid (pausi kaugus lõigu, lause ja fraasi algusest ning samuti kaugus eelnevast pausist ning eelnevast hingamisest).

Prognoositavaks tunnuseks oli pausi kestus, lineaarse regressiooni tarvis tuli funktsioonitunnus logaritmida, sest logaritmitud kestus allub normaaljaotusele rohkem. Pauside kestuste modelleerimise varasemad katsetused on näidanud, et sõnadevaheliste pauside kestused on mitmesel regressioonil prognoositavad.[14] Selles töös on rakendatud erinevaid prognoosimeetodeid: klassikalist regressioonanalüüsi, klassifikatsiooni ja regressioonipuid (CART-meetodit) ning närvivõrke. Modelleerimisel kasutati statistikaprogramme SYSTAT 11 ja SAS 9.1.

Pauside kestuste mitmese regressiooni abil modelleerimisel osutusid tekstistruktuuri tunnustest olulisteks lõigu-, lause- ja fraasilõpp. Pausile eelneva kõnetakti tunnustest osutus oluliseks usaldusnivool 0,05 vaid binaarne tunnus, mis näitas, kas pausile eelnev kõnetakt oli pikendatud või mitte. Lisaks oli kestuse prognoosimisel oluline konkreetse pausi kaugus talle eelnevast pausist. Pausi kestuse logaritmiline arvutusvalem on järgmine:[15]

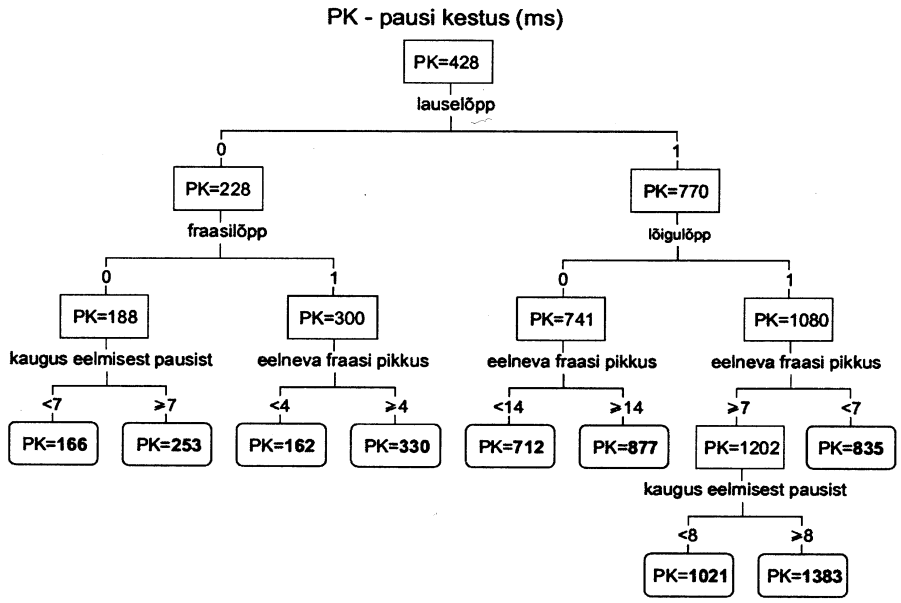
$$\textit{LN (pausi kestus)} = -1{,}973 + 0{,}373 \times LQLQP + 1{,}454 \times LALQP + 0{,}441 \times FRKOM + 0{,}012 \times KAUGFR + 0{,}024 \times KAUGPA + 0{,}133 \times PIKENDUS$$

Pausi kestuse erinevate meetoditega modelleerimisel jäi prognoosiviga 29–37% piiresse. Arvestades pauside suurt variatiivsust ja seda, et algmaterjali moodustasid 26 diktori kõneandmed, siis oli selline suur veaprotsent paratamatu. Väikseimad veaprotsendid olid mitmese regressioonanalüüsi ja närvivõrkude meetodi rakendamisel, suurim prognoosiviga tuli CART-meetodil. Samal ajal on tulemused kõige paremini interpreteeritavad just

[13] E. Zvonik, F. Cummins, Pause Duration and Variability in Read Texts. – Proceedings of the 7th International Conference on Spoken Language Processing. ICSLP-2002. Denver, 2002, lk 1109–1112.

[14] M. Mihkla, Modelling Pauses and Boundary Lengthenings in Synthetic Speech. The Second Baltic Conference on Human Language Technologies. Tallinn, April 4–5. 2005. Tallinn, 2005, lk 305–310.

[15] Muutujad valemis on: LQLQP – lõigulõpu tunnus, LALQP – lauselõpu tunnus, FRKOM – fraasilõpp (koma), KAUGFR – eelneva fraasi pikkus, KAUGPA – kaugus eelmisest pausist, PIKENDUS – viimase kõnetakti pikendus.



Joonis 3. Teksti pauside kestuste modelleerimise regressioonipuu. Ümmarguste nurkadega kastid kujutavad regressioonipuu lehti, kus on kirjas vastava pausiklassi kestus.

regressioonipuu abil. Joonisel 3 on kujutatud kahendpuu, mis aitab määrata pauside kestust konteksti põhjal.

Regressioonipuu abil toimub pauside esmane liigitus selle põhjal, kas on tegemist lauselõpu pausiga või mitte. Vasakusse harusse satuvad lausesisesed pausid ja paremasse lauselõpu pausid. Lausesisesed pausid jagunevad omakorda kaheks selle põhjal, kas on tegu fraasilõpuga või mitte. Fraasisisese pausi pikkuseks on 166 millisekundit, kui eelnev paus oli vähem kui seitse kõnetakti tagasi, vastasel juhul on fraasisisese pausi kestuseks 253 millisekundit. Kahendpuult on näha, et fraasilõpu paus võib olla fraasisisesest pausist pisut lühem (162 ms versus 166 ms). Fraasilõpu pausi sellesse harusse satuvad eelkõige loetelu elemendid, mille korral fraasi pikkus piirdub 1–3 kõnetakti või sõnaga. Analoogiliselt on seletatavad regressioonipuu teised harud, kus fikseeritakse fraasi-, lause- ja lõigulõpu pauside kestused. Erinevalt regressioonvõrrandist puudub kahendpuust tunnus, mis näitab pausieelset lõpupikendust. Sisemistest parameetritest sõltuvalt võib regressioonipuid genereerida suurema või väiksema harude arvuga ja erineva lehtede arvuga. Suuremaharuliste puude puhul on ületreenituse oht, s.t nad kirjeldavad väga hästi konkreetset andmekogumit, aga uute sisendandmete puhul võib mõni haru anda eksitavaid tulemusi.

Kõik pauside kestusi kirjeldavad mudelid olid statistiliselt olulised ja kirjeldasid meetodist sõltuvalt 66–71% funktsioonitunnuse variatiivsusest.

Lõpupikenduste modelleerimine ebaõnnestus, sest saadud mudelid kirjeldasid vaid 25–30% pikenduste muutumisest.

Pauside asukoha prognoosimiseks rakendati esmalt logistilist regressiooni, millega prognoositi tõenäosust, kas mingi sõna järel kõnevoos tehakse paus või mitte. Logistilise regressiooni muutujatena kasutati samu tunnuseid mis pauside kestuste ennustamisel. Peale nende kaasati veel kaks binaarset tunnust, mis näitasid, kas järgnev sõna on pärisnimi või võõrsõna. Neid tunnuseid ärgitas lisama kujutelm, et pärisnimede ees (nt *Minu nimi on Tamm, Jüri Tamm*) ja võib-olla ka enne keerulisemate võõrsõnade väljaütlemist (nt *Rahvas toetas konstitutsioonilist monarhiat*) tehakse kõnes väikene paus. Etteruttavalt võib öelda, et see hüpotees ei leidnud tõestust. Pärisnimedel ja võõrsõnadel oli pausiga väga nõrk korrelatsioon, need tunnused osutusid ebaolulisteks.

Tabel 3.

Logistilise mudeli abil ennustatud pauside asukohad

	Mudel õigesti ennustanud	Pauside tegelik arv kõnes	Korrektsuse protsent
Paus sõna järel (PAUS = 1)	402	600	67
Pausi ei ole (PAUS = 0)	2510	2708	93
KOKKU			88

Tabelist 3 on näha, et mudel on ennustanud õigesti 67% pausi asukohtadest. Kogu mudeli prognoositäpsus, s.t hinnang, kas mingi sõna järel on paus või mitte, küünib 88 protsendini Analüüsi põhjal tegime kindlaks, et mõningate tekstistruktuuri märgendite (lõigulõpu, lauselõpu, kooloni, mõttekriipsu) järel on paus 93–100% tõenäosusega. Kui sellised kindlad pausid mudelist välja jätta, siis suudab see lausesiseste pauside mudel prognoosida lausesiseseid pause vaid 44-protsendilise täpsusega.

Tabelis 4 on logistilise regressiooni abil leitud kuus tunnust, mis mõjutavad lausesisese pausi asukohta. Väga oluliselt tingib lausesisest pausi kõnevoo tekstis olev koma, pausi võimalikkus on sel juhul 17,4 korda keskmisest

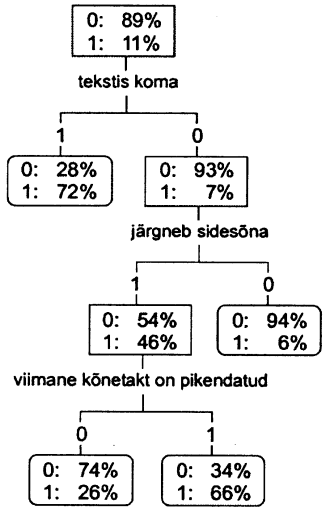
Tabel 4.

Logistilise regressiooni tulemused: lausesisese pausi asukohta oluliselt mõjutavad muutujad, nende võimalikkuse suhe ja usalduspiirid

Sõltumatud muutujad	Võimalikkuse suhe	Usalduspiirid	
		Alumine	Ülemine
Sõna järel on tekstis koma	17,4	11,7	25,9
Järgmine sõna on sidesõna	7,9	4,8	12,8
Sõna kaugus lause algusest	1,1	1,0	1,2
Viimase kõnetakti pikkus	1,3	1,1	1,4
Viimase kõnetakti välde	1,2	1,1	1,5
Viimane kõnetakt on pikendatud	6,9	5,2	9,2

suurem. 7–8 korda suurendab sõnajärgse pausi võimalikkust see, kui järgnev sõna on sidesõna või kui selle sõna kõnetakt on pikendatud. Keskmisest pisut sagedamini tehakse kõnes paus pikemate kõnetaktide ja pikemavälteliste sõnade järel. Nende tunnuste osa kipub siiski prognoosimudelis marginaalseks jääma, sest nad tõstavad pausi esinemise võimalikkust vaid 1,2–1,3 korda.

Joonisel 4 esitatud regressioonipuu võtabki lihtsustatult kokku logistilise regressiooni tulemused: suure tõenäosusega tehakse lause sees kõnevoos paus komaga sõna järel ja vaid sellise sidesõna ees, millele eelnenud sõna kõnetakt oli pikendatud.



Joonis 4. Lausesisese pausi asukoha prognoosimise regressioonipuu.

Kokkuvõte

Sidusa kõne analüüs näitas, et eri liiki pausid on kõnes kestuselt eristatavad. Pauside kestus ja nende asukoht kõnevoos on modelleeritav statistiliste meetoditega. Pauside suurest variatiivsusest johtuvalt, kirjeldavad saadud mudelid eri diktorite keskmistatud hääleparameetreid ning pauside kestuste ja nende paiknemise kõige üldisemaid seaduspärasusi. Seega peaks paari diktori kohta koguma mahuka kõnematerjali ja sama metoodikat rakendama ühe diktori andmestikul eraldi. Lõpupikenduste prognoosimiseks ei õnnestunud arvestatavat kestusmudelit luua. Ilmselt tuleb lõpupikendusi käsitleda häälikute kestusmudeli osana. Töös esitatu on üks samme, mis aitab prosoodia reeglipõhiselt genereerimiselt üle minna statistiliste mudelite kasutamisele kõnesünteesil.

VI

Mihkla, Meelis 2006. Comparision of statistical methods used
to predict segmental durations. – The Phonetics Symposium 2006:
Fonetiikan Päivät 2006, Helsingi, 30.–31.08.2006. (Toim.)
Aulanko, Reijo; Wahlberg, Leena; Vainio, Martti.
Helsingi: 120–124, University of Helsinki.

COMPARISON OF STATISTICAL METHODS USED TO PREDICT SEGMENTAL DURATIONS*

"Does the Colour of Cat Matter - if It Catches Mice?"

Oriental proverb

Meelis Mihkla
Institute of the Estonian Language, Tallinn, Estonia
meelis@eki.ee

Abstract

Different techniques (linear regression, CART, neural networks) were used to predict vowel segmental durations. The input consisted of the durations of sounds and pauses, measured from the speech of two radio newsreaders, and certain features generated from readout text. The choice and values of the features were kept similar for all methods to be compared. The software used was SAS 9.1. The methods were evaluated for their predictive error, model interpretability, necessity for preliminary data processing, and some other criteria. Somewhat surprisingly, the predictive precision of linear regression turned out to be the same, if not better than that of the nonlinear methods. As for CART, the predictive error was slightly larger, but the interpretability of the corresponding model was the best.

Keywords: segmental duration, prediction, linear regression, neural networks, regression trees

1 Introduction

What is a *good* method to predict speech prosody? Are there any objective criteria for choosing the *best* statistical technique? These are some of the questions faced by anyone trying to model speech prosody from fluent speech by statistical methods. The first personal impulse pushing me to cogitate about those problems happened to be the plenary talk given by Yoshinori Sagisaka to the Congress of Phonetic Sciences in Barcelona, 2003, containing a statement to the effect that with their over 20-years experience in modelling speech prosody they prefer regression analysis (Sagisaka 2003). Reading literature in the field (e.g. Brinckmann 2004, Campbell 2000, Horak 2005, Sreenivasa & Yegnanarayana 2004, Vainio 2001), one will notice that in most cases prediction of speech prosody is done using neural networks or regression trees rather than regression analysis. Reasons for the choice of this or that method, however,

* The paper was supported by the state programme "Language technological support for the Estonian language"

are hardly ever given, while the results are usually compared to those of a rule-based prosody generator. Thus it seems that each concrete case of method choice is determined rather pragmatically, depending on the educational background of the researcher, the influences of his or her tutors or colleagues, availability of software, etc.

The aim of the present paper is to test the goodness of certain techniques (regression, CART, neural networks) as methods of predicting sound segmental duration, comparing their performance on same data. The evaluative criteria include predictive error, result interpretability, necessity for preliminary data processing etc.

2 Initial data and argument features

The initial data consisted of fragments and full texts of news read out by two radio newsreaders (one male, one female). In total, 6.3 minutes of male and 9.1 minutes of female speech was analysed, being segmented into 5063 and 7010 sounds, respectively.

Our choice of argument features was based on the duration rules developed by Klatt (Klatt 1979) and the experience of other researchers (Vainio 2001, Horak 2005) in predicting sound durations was also considered. The main factors affecting the duration of a sound include the phoneme environment of the sound, its position in the syllable, the position of the syllable in the word, the position of the word in the phrase etc. Naturally, such special Estonian features as its three-way system of phonetic quantity and certain foot-bound features could not be overlooked either. In total, 26 argument features were generated from the text. To optimize the number of features some preliminary analysis was carried out. By means of linear regression it was found out which features were significant for modelling both male and female speech. The number of such features was 18 (see Table 1).

Table 1. Inputs (per sound) for modelling segmental durations.

	Inputs (per sound)	*Measurement*
1.	before last phoneme class	nominal (9 classes)
2.	before last phoneme length	Binary
3.	previous phoneme class	Nominal
4.	previous phoneme length	Binary
5.	current phoneme identity	nominal (26 phonemes)
6.	current phoneme length	Binary
7.	next phoneme class	Nominal
8.	next phoneme length	Binary
9.	next but one phoneme class	Nominal
10.	next but one phoneme length	Binary
11.	phoneme position in syllable	Ordinal
12.	syllable position in foot	Ordinal
13.	length of word in feet	Ordinal
14.	monosyllabic word	Binary
15.	length of phrase in words	Interval
16.	length of sentence in phrases	Interval
17.	final word of phrase	Binary
18.	final word of sentence	Binary

For all three methods the response to be analysed consisted of logarithmic

durations of sounds. True, neither the neural networks nor CART require a Gaussian distribution, but some normalization would certainly do no harm to neural networks.

3 Statistical analysis

Statistical modelling of sound durations was done by means of the program package SAS 9.1 (see Fig.1 for the block diagram).

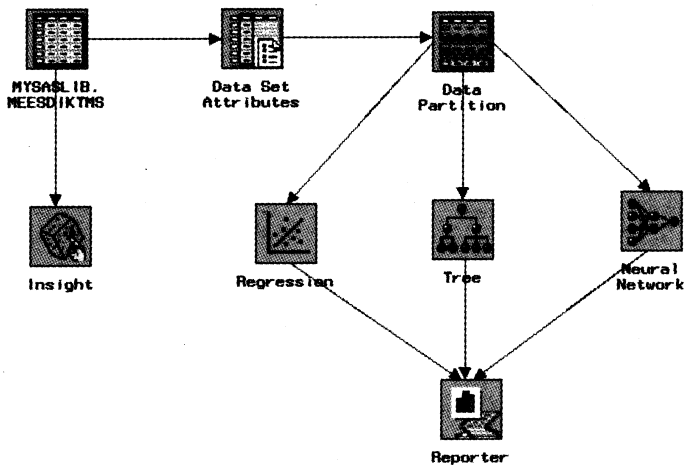


Figure 1. SAS Enterprise Miner workspace.

Most of the modelling was done on the basis of the default parameters of SAS 9.1. Table 2 contains the general model parameters. For modelling purposes each newsreader's data were divided into three sets: 50% of the data was used for model training, while 30% was kept for data validation and 20% for testing.

Table 2. Inner settings of prediction models.

Prediction methods	*Inner settings of models*
Neural network	Multilayer perceptron One hidden layer Selection criteria – average error
Regression	Linear regression Method – backward/none Significance level – 0.05
Classification and regression trees	Splitting criterion – F-test Significance level – 0.2 Maximum depth of tree – 6

4 Results

For the results of statistical modelling see Table 3. Notably, on same data and same argument features all three methods yielded a very similar error percentage.

Surprisingly enough, the lowest error percentage belonged to linear regression. As we know, linear regression has been expected to reveal only the clearest and most direct correlation between the input and output. And even though some nonlinearity is known to be smuggled into the regression model by logarithming of the response and nonlinear coding of the input features anyway, it is believed that the more covert correlations between input and output can be reached only by such more complicated nonlinear methods as classification and regression trees (CART) and neural networks. However, as is proved by our results, linear regression - with its long history of success in speech wave processing (Markel & Gray 1976) and continuing popularity in speech analysis and synthesis – should not be overlooked in modelling the temporal structure of speech.

As for clarity of interpretation, the best method is the binary tree, while regression coefficients also seem to be mirroring the influence of input on sound durations quite logically. It is far more difficult, however, to interpret the results of the training process on neural networks. Although of the three methods under discussion, linear regression is the only one directly requiring a Gaussian distribution of the response, normalisation improves the stability of the neural networks model. Before statistical modelling can be started, both regression analysis and neural networks require some preliminary processing of the argument features. For regression analysis the nominal features need to be replaced by a large number of binary pseudo-features. For neural networks the input variation zone has to be transformed to the interval [0, 1]. The last three features in Table 3 are indirect criteria for method evaluation, while the very last criterion rather refers to a virtue of the SAS statistics programme – C score code generation for the rather complicated neural NW models.

Table 3. Prediction errors and features of predictive modelling nodes.

Features		*Neural NW*	*Regression*	*CART*
Prediction errors:	Training	0.230	0.230	0.250
- male newsreader	Validatio	0.243	0.248	0.264
(average absolute error 21%)	Testing	0.230	0.232	0.255
- female newsreader	Training	0.224	0.221	0.230
(average absolute error 19%)	Validatio	0.221	0.218	0.231
	Testing	0.221	0.217	0.230
Model interpretation		complicated	good	very good
Response normalisation		dispensable	necessary	unnecessary
Processing inputs		necessary	necessary	unnecessary
Interactive training		yes	no	yes
Model with missing inputs		no	no	yes
C score code		yes	no	no

5 Conclusion

It was surprising to find that with linear regression, predictive precision was the same if not better than with the nonlinear methods. As for CART, it had a little higher predictive error, but then the model excelled in interpretability; while in addition, there is no need either for preliminary input processing or for response normalisation. And

yet, all three models could do with some inner tuning, while the proof of the actual goodness of any method is in the practice, i.e. in perception tests of synthetic speech. Thus we have to confess that this time we did not really succeed in pinpointing the „best" predictive method. Each method has its good points as well as drawbacks and so the truth to be followed lies in the oriental proverb serving as the motto to the present paper: „*Does the colour of cat matter - if it catches mice?*".

6 References

Brinckmann, Caren (2005). The "Kiel corpus of read speech" as a resource for prosody prediction in speech synthesis. In M. Langemets & P. Penjam (Eds.), *Proceedings of the second Baltic conference on Human Language Technologies* (pp. 101-106). Tallinn: Institute of Cybernetics, Institute of Estonian Language.

Campbell, Nick (2000). Timing in speech: a multilevel process. In M. Horne (editor), *Prosody: theory and experiment* (pp. 281-334). Dordrecht/Boston/London: Kluwer Academic Publishers.

Horak, P. (2005). Using neural networks to model Czech text-to-speech synthesis. In R. Vich (editor), *Proceedings of the 16th Conference of electronic speech signal processing* (pp. 76-83), Prague: TUDpress.

Klatt, D. H. (1979). Synthesis by rule of segmental durations in English sentences. In B. Lindblom & S. Öhman (eds.), Frontiers of Speech Communication research (pp. 287-300), New York: Academic Press.

Markel J. D. & Gray A. H. (1976). Linear Prediction of Speech. Berlin/Heidelberg/New York: Springer-Verlag..

Sagisaka, Yoshinori (2003). Modeling and perception of temporal characteristics in speech. In M. J. Sole, D. Recasens & J. Romero (eds.), *Proceedings of 15th International Congress of Phonetic Sciences* (pp. 1-6). Barcelona, 2003.

Sreenivasa Rao, K., Yegnanarayana, B. (2004) Modeling syllable duration in Indian languages using neural networks (pp. 313-316). In *Proc. Int. Conf. Acoust., Speech Signal Processing*, Montreal, Quebec, Canada.

Vainio, Martti (2001). Artificial neural network based prosody models for Finnish text-to-speech synthesis. Helsinki: University of Helsinki.

VII

Mihkla, Meelis 2007. Morphological and syntactic factors
in predicting segmental durations for Estonian text-to-speech synthesis. –
Proceedings of the 16th International Congress of Phonetic Sciences,
Saarbrücken, 6–10 August 2007, (Toim.) Jürgen Trouvain,
William J. Barry. Saarbrücken: 2209–2212.

MORPHOLOGICAL AND SYNTACTIC FACTORS IN PREDICTING SEGMENTAL DURATIONS FOR ESTONIAN TEXT-TO-SPEECH SYNTHESIS

Meelis Mihkla

Institute of the Estonian Language
Roosikrantsi 6, Tallinn, ESTONIA
meelis@eki.ee

ABSTRACT

Traditionally, durational models of speech units have been developed without paying much heed to morphology and part-of-speech information while predicting speech temporal structure. The aim of the present study was to find out whether the rich morphology of the Estonian language could possibly provide some additional (beside the syntactic and part-of-speech) information that could be used in predicting durations. The project is a continuation of prosody studies for Estonian text-to-speech synthesis. Sound durations in the speech of radio newsreaders were modelled by means of different statistical methods (linear regression and neural networks). Model input consisted not only of descriptors of sound context and position, but also of information on part of speech, part of sentence and morphological features. The results indicated a decrease of error in the prediction of segmental durations. Such results were in good harmony with our expectations concerning a morphologically rich language.

Keywords: morphological factors, part-of-speech, segmental durations, TTS synthesis

1. INTRODUCTION

Speech prosody being affected by very many factors and their complicated combined effects it is not easy to generate synthetic speech with a prosodically appropriate temporal structure. As a rule, morphological features are not included among the factors relevant for the temporal structure of speech (cf. [1], [5] and [6]). One reason may be that most of the studies hitherto available on text-to-speech synthesis concern languages with relatively little morphology. Finnish is different in that respect, and they, indeed, have a study on the role of morphological features on the duration of speech units [7]. As Estonian is a language with the word having a central role in grammar as well as phonetics, and with an extremely rich morphology at that, we wondered if the temporal structure of Estonian speech could possibly be affected by some morphological, lexical and maybe even syntactic features.

In some previous studies on Estonian prosody several statistical methods (linear regression, neural networks, CART) were successfully applied to predict the segmental duration of speech units [2], [3]. In the present paper the same methods were extended over certain linguistically based factors such as morphology, lexis and syntax. The linguistic knowledge used rests on available technologies prepared for the Estonian language [4], [8]. As the morphological analyser and parser already work in text-to-speech synthesis it seemed a waste not to use their information in our prosody generator. A demo version of the Estonian syntax analyser is already accessible over the Internet http://www.cs.ut.ee/~kaili/parser/demo/.

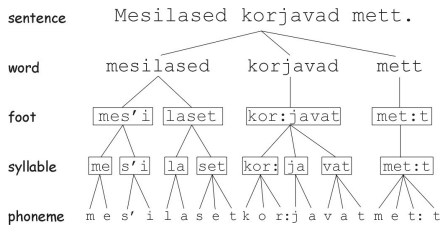
The most natural way to assess the effect of morphological, part-of-speech and syntactic factors seemed to lie through an extension of our previous methodology of statistical modelling to see how those factors may affect the functioning of durational models. The models were built on two different methods - linear regression and neural networks, the latter being a nonlinear method. The comparable effect of the factors was measured by change of output error as measured on different models.

2. INITIAL DATA AND ARGUMENT FEATURES

The initial data consisted of speech passages and news texts as pronounced by two radio newsreaders (male and female). The total amount of speech analysed included 6.3 minutes of male speech and 9.1 minutes of female speech

segmented manually into 5063 and 7010 speech sounds, respectively.

Figure 1: Hierarchical encoding of the relative position and length of a current speech unit. In this case for example the phone [l] is being estimated and its place is coded according to its position in syllable [la] of length two (phones). The syllable's position is coded in relation of foot [lased] with a length of two syllables. The foot's position is coded in relation to word [mesilased] with a length of two feet. The word is further given a code according to its place in the sentence [Mesilased korjavad mett. *'Bees gather honey.'*] with the length three words.



In durational models of Estonian sounds the argument features are represented hierarchically (Figure 1). The hierarchical levels are sentence, word, foot, syllable and phoneme. So the relative position of a speech unit, e.g. phoneme, is referred to in the hierarchical scale by describing its position in the syllable, the position of the syllable in the foot, the position of the foot in the word, and, finally, the position of the word in the sentence. In addition, as has been proved by previous studies, information on sentence and word length comes in handy.

Table 1: Input (per sound) for modelling segmental durations.

Input	*Measurement*
Left phoneme class	Nominal (9 classes)
Left phoneme length	Binary (short and long)
Current phoneme identity	Nominal (26 phonemes)
Current phoneme length	Binary
Right phoneme class	Nominal
Right phoneme length	Binary
Phoneme position in syllable	Ordinal
Syllable position in foot	Ordinal
Length of word in feet	Ordinal
Length of phrase in words	Interval
Punctuation	Binary

The phoneme segment itself is characterized by phoneme identity and phoneme length. The necessary characteristics also include the class and length of the left and right neighbours (predecessor and successor) of the current phoneme. Before supplying the morphological, part-of-speech and syntactic features the durational models were optimized, removing all but the most vital features. The aim was to reduce not only the number of features to be considered but also to avoid some possible joint effects between new and old features (see Table 1 for the features proved the most essential by our analysis).

Table 2 represents some new candidates of argument features for the input of durational models. Most of the possible features are contained in the morphological factor. The values of the morphological features and the part-of-speech information have been generated by means of the Estonian morphological analyser [8]. As can be seen, most of the morphological factors concern a certain part of speech only. Verbs, for example, are involved with the highest number of factors, whereas adverbs, adpositions and conjunctions carry a single morphological marker - invariable word. For nearly all new factors the influence is manifested on word level, except the feature "stem vs. suffix", that belongs to phoneme level. That feature was included to check whether the duration of stem sounds might differ from that of suffix sounds. The syntactic analysis of the text sentences was done manually. The response of the models consisted of logarithmed durations of sounds.

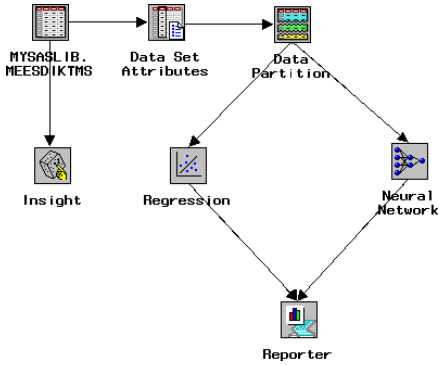
Table 2: Morphological, syntactic and lexical factors concerning the word. Number of the necessary values encoded as input to duration models.

Input	*Number of stages*
Morphological factors:	
- case	15
- number	2
- tense	2
- person	2
- voice	2
- infinit/partic	6
- stem/suffix	2
Syntactic factor:	
- part of sentence	8
Lexical factor:	
- part of speech	11

3. STATISTICAL ANALYSIS

Statistical modelling of the sounds was done by means of the SAS package 9.1. Figure 2 contains the block diagram of our data processing.

Figure 2: SAS Enterprise Miner workspace.



The methods picked were linear regression and the neural networks. The choice was purely pragmatic, enabling a comparison of the response of a linear and a non-linear model to different argument features. Linear regression used backward selection with a 0.05 significance level. The neural networks represented a multilayer perceptron with one hidden layer. For modelling purposes each newsreader's data were divided into three parts: 50% of the data were used for model training, 30% for data validation and 20% for testing.

4. RESULTS

Table 3 demonstrates a decrease of the average error in response to different information added to model input. The results indicate that, depending on the factor, part-of-speech and part-of-sentence information as well as morphological features can improve model efficiency and predictive precision by ca 0.4 – 1.3 %. Without input of morphological-syntactic information the average predictive error of segmental durations stayed within the limits of 16.5 – 18.1 %. Of morphological features separate mention has been made of stem vs. suffix. As can be seen from the table the output effect of that feature is irrelevant. Consequently, the duration of suffix sounds does not differ from that of stem sounds. The final row of Table 3 gives the total contribution of all factors to the efficiency increase of the durational model. A can be seen, the total decrease of the predictive error does not equal the sum of the effects of the individual factors, which is obviously due to co-effects occurring between some factors. The sensitivity of neural networks is raised considerably by supplying the input with the part-of-sentence (syntactic) feature. The error decreases twice as much as in the regression model. Nevertheless it is rather difficult to express the error decrease in quantitative terms. The final assessment of the effect of morphological-syntactic input is awaiting some perception tests.

Tabel 3: Results of adding morphological information, part-of-speech and part-of-sentence status to the durational models. The values represent the average decrease in error (%).

Factors	Male newsreader		Female newsreader	
	REGR	NN	REGR	NN
Part of speech	-1,02	-0,46	-1,34	-1,09
Morphology	-0,96	-0,71	-0,84	-0,86
Part of sentence	-0,37	-0,82	-0,46	-1,06
Stem vs. suffix	-0,04	-0,08	0,00	-0,03
All factors together	-1,64	-1,29	-1,61	-2,36

Table 4: The values of regression coefficients for different part-of-speech in the male and female material.

Part of speech	Male newsreader	Female newsreader
Proper noun	6.23	5.22
Noun	2.25	2.10
Adposition	0.82	2.82
Genitive attribute	0.42	1.35
Verb	0.00	0.00
Numeral	-0.10	0.42
Conjunction	-0.14	1.81
Adjective	-0.39	1.14
Adverb	-0.89	-2.90
Pronoun	-4.13	-3.86
Ordinal numeral	-5.44	-7.48

A far as the models analyzed are based on the speech of no more than two speakers it is certainly premature to make generalizations. Yet a visual survey of the regression coefficients suggests that the most distinct regularities concern the parameters of the part-of-speech factor. Table 4 represents the values of the regression coefficients for different parts of speech in the male and female material. Variance seems to be higher in the middle part of the table, whereas the beginning and end parts are very similar. The table reveals that the speech sounds of proper names are pronounced in average 5.22 – 6.23 ms (10 %) longer than in verbs. The average duration of sounds in newsreaders speech was 62.5 and 64.1 ms,

respectively. Nouns and adpositions were slightly prolonged. It was surprising to find such lengthening in adpositions as in most languages function words are shorter than content words. An Estonian adposition invariably belongs to a noun phrase. The noun often stands in the focus of the sentence, while its more than average length may extend to a neighbouring adposition. Ordinal numbers, however, were pronounced over 10% shorter, while pronouns and adverbs tended to be shorter by ca 5%.

5. CONCLUSION

The study was meant first and foremost as a continuation of prosody research for Estonian text-to-speech synthesis. The results revealed that addition of morphological-syntactic information to model input yields a couple of percent decrease of error in predicting segmental durations. Considering the rich morphology of the Estonian language such behaviour of the models was no surprise. As a morphological analyser is already at work in the linguistic processing of texts for Estonian speech synthesis it is obvious that some part-of-speech and morphological information will be used in duration modelling. The possible necessity of involving the rather complex syntactic analyser, however, cannot be decided without first performing some perception tests.

6. REFERENCES

[1] Campell, N. 2000. Timing in speech: a multilevel process. In M. Horne (editor), *Prosody: theory and experiment*. Dordrecht/Boston/London: Kluwer Academic Publishers, 281-334.

[2] Fishel, M., Mihkla, M. 2006. Modelling the temporal structure of newsreaders' speech on neural networks for Estonian text-to-speech synthesis. In: *Proceedings of the 11th International Conference "Speech and Computer": SPECOM2006*, St. Petersburg: Anatolya Publishers, 303 - 306.

[3] Mihkla, M., Kuusik, J. 2005. Analysis and modelling of temporal characteristics of speech for Estonian text-to-speech synthesis. *Linguistica Uralica*, XLI(2), 91 - 97.

[4] Müürisep, K. 2000. Eesti keele arvutigrammatika: süntaks. Dissertationes Mathematicae Universitatis Tartuensis 22.

[5] Sagisaka, Y. 2003. Modeling and perception of temporal characteristics in speech. In M. J. Sole, D. Recasens & J. Romero (eds.), *Proceedings of 15th International Congress of Phonetic Sciences*. Barcelona, 1–6.

[6] van Santen, J. 1998. Timing. In: *Multilingual text-to-speech synthesis*: The Bell Labs Approach, Sproat, R. (editor), Kluwer Academic Publishers, 115-140.

[7] Vainio, M. 2001. Artificial neural network based prosody models for Finnish text-to-speech synthesis. Helsinki: University of Helsinki.

[8] Viks, Ü. 2000. Eesti keele avatud morfoloogiamudel. -- *Arvutuslingvistikalt inimesele* (toim T. Hennoste). Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1. Tartu, 9-36.

7. ACKNOWLEDGEMENT

The study was financed by the national programme “Language technological support of Estonian”.

VIII

Mihkla, Meelis 2007. Modelling speech temporal structure for Estonian text-to-speech synthesis: feature selection. Trames. Journal of the Humanities and Social Sciences, 11(3): 284–298.

TRAMES, 2007, **11**(**61**/**56**), 2, 284–298

MODELLING SPEECH TEMPORAL STRUCTURE FOR ESTONIAN TEXT-TO-SPEECH SYNTHESIS: FEATURE SELECTION

Meelis Mihkla

Institute of the Estonian Language, Tallinn

Abstract. The article discusses the principles of selecting features for modelling the temporal structure of Estonian speech, using different types of read-out texts, with a view to text-to-speech synthesis (TTS). Feature selection is known to depend on certain general issues regulating speech temporal structure, as well as on some language specific aspects. The durational model of Estonian stands out for some foot-bound features (foot quantity degree, number of feet in the word) being included in the input. In addition to the traditional descriptors of sound context and hierarchical position the prediction of Estonian segmental durations requires information on some morphological, syntactic and lexical features of the word, such as word form, part of sentence, and part of speech. In the prediction of pauses in the speech flow the relevant features are: distance from sentence beginning and from the previous pause, the length and quantity degree of the preceding foot, and the occurrence of a punctuation mark or conjunction. Although expert opinions were used in feature selection, statistical methods should be applied to test the vector of optimal argument features.

Keywords: feature selection, speech timing, segmental durations, pauses, text-to-speech synthesis, feature significance, statistical modelling

1. Introduction

Variability is one of the keywords of speech technology. While in speech recognition variability in the speech wave is a frequent source of trouble, insufficient variability in speech synthesis may lead to monotony and unnaturalness of synthetic speech (Tatham, Morton 2005:9). Synthetic speech cannot boast a natural temporal structure without successfully rendering the normal variability in the duration of sounds and pauses, as well as in the positioning of pauses in the speech flow. This is a complex task requiring an optimal choice of features to model speech timing. Modelling speech temporal structure is of particular interest for speech technology, representing an interface between the cognitive and mechanical aspects of speech

generation. The general aspects governing that structure derive from the human vocal tract and articulatory mechanisms being the same for all languages. The specific aspects, however, are connected with factors typical of the language and the speaker. It is, for example, highly probable that one and the same sound sequence as pronounced by two different speakers (or even by the same speaker on two different occasions) has different timing characteristics (Campbell 2000:281).

Generation of synthetic speech with a prosodically appropriate temporal structure is never easy as speech prosody is subject to the influence of many factors, often with complex joint effects. Moreover, a language may bring forth a lot of factorial coincidence resulting in a surprising number of exceptional cases (van Santen 1998:115). Characteristics controlling speech timing have been studied for a long time and in several different speech-related spheres. However, not all discoveries have as yet been circulated in full, neither have all factors been integrated in a single extensive model to be used in all relevant spheres. Even confining oneself to just one concrete sphere, it is hard to comprehend the underlying principles and mechanisms of the characteristics (Sagisaka 2003:1).

Over the past decade the developments in speech synthesis have revealed a certain tendency of unification and multilingualism, which means that similar development systems, methods and approaches are applied to many different languages. The existing Estonian text-to-speech synthesizer, for example, uses the MBROLA synthesis engine worked out at Mons University, Belgium (Dutoit 1997:276). The question remains, however, to what extent unified algorithms and unified feature selection could be applied in speech prosody, and in particular, in the modelling of speech timing.

Characteristics of segmental durations have been measured for many languages in order to pinpoint the universal and the specific in the timing patterns of languages (Sagisaka 2003:1). Speech technology has been striving to obtain precise control over segmental durations in order to synthesize speech with a natural-sounding rhythm and timing. The present paper is focused on the selection of optimal characteristics to model speech temporal structure for Estonian text-to-speech synthesis.

2. Principles of feature selection for modelling timing in speech

There have been three approaches in the control of speech timing: one is mora-timed rhythm, which is used, e.g., in Japanese, the second is syllable-timed rhythm – every pronounced syllable is supposed to take up roughly the same amount of time, and the third is stress-timed rhythm, recognized and used in many Germanic languages.

In Japanese, mora isochrony has been observed as a temporal constraint controlling vowel duration. A negative correlation has been found to exist between the durations of vowels and their neighbouring consonants. The fact that the temporal compensation of the duration of a vowel is more influenced by the duration of its preceding consonant is regarded as an acoustic manifestation of mora-timing. As has

been proved by statistical analysis, such compensation takes place in mora units, not in syllables (Sagisaka 2003:2). This does not of course exclude more extensive regulation. Speech rhythm is readjusted on phrase level - the more moras in the phrase, the shorter their average durations. In the end it is the mora constraints and the local adjustment of speech rate within phrases that determine most of the variation of segmental durations in read-out Japanese speech (Sagisaka 2003:2). Mora metrics has been applied quite successfully in Estonian phonology as well. In Estonian word prosody Arvo Eek has interpreted intra-foot quantity as a manifestation of mora isochrony, where the quantity degree is determined by the distribution of durations within the foot (Eek, Meister 2004:336–357).

In a syllable-timed language, every syllable is thought to take up roughly the same amount of time when pronounced, though the actual duration of a syllable depends on the situation. Spanish and French are commonly quoted as examples of syllable-timed languages. When a speaker repeats the same sentence many times at the same rate of articulation, the durations of adjacent phones display a strong negative correlation, i.e. any variance in the duration of a single phone is compensated in the adjacent phones and so the temporal sequence of articulation must be organised at levels higher than phoneme, e.g. syllable (Huggins 1968). The syllable-timing hypothesis was proposed by Campbell and Isard in statistical modelling to account for the interaction between higher and lower levels of timing control (Campbell and Isard 1991). It posits the syllable as a mediator and offers a way to map the effects of linguistic and semantic contexts onto the physiological constraints of speech sound production. By adopting a higher level framework for duration control, it overcomes the sparcity of data problem in the modelling of the variability in individual phone durations (Campbell 2000:307).

In a stress-timed language, syllables may last different amounts of time, but there is a constant amount of time (on average) between two consecutive stressed syllables. For English the rules of speech timing were formulated by Dennis Klatt. Using the results of other researchers he made up 11 rules describing 84% of segment temporal variation in a text read out by himself (Klatt 1979). Klatt's rule-based model has been modified and developed by other researchers. Jan van Santen generalized the Klatt rules, adding the following six factors affecting segmental durations in American English (van Santen 1998:123–124):

1. phonetic segment identity (30 values),
2. identities of surrounding segments (10),
3. syllabic stress (3),
4. word 'importance' (contrastive stress),
5. location of the syllable in the word,
6. location of the syllable in the phrase

Thus the above models are centred on the stress groups (syllable stress, contrastive stress). Rule-based models were good enough to provide reasonable segment durations for most cases, yet sometimes grave mistakes occurred. The mistakes were often due to attempts of simultaneous application of independently derived rules. The advent of large databases, however, enabled the use of statistical

methods to predict segmental durations much more precisely. Many statistical models have made sound use of the parameters and features of Klatt's rule-based model. Those have been applied as argument features, either directly or selectively, integrating some language-specific information. Horák, for example, introduced a special feature of monosyllabic words in his model of Czech durations (Horak 2005:79), and Vainio roped in some morphological features and part-of-speech information while modelling prosody for Finnish TTS (Vainio 2001:66–67). Timing control is just an aspect depending on language-specific features.

3. Feature selection for Estonian TTS

Chapter 2 described three basically different phonological approaches to feature selection for a durational model: one was based on mora metrics, the second on syllables and the third on stress-timing. Although certain characteristics allow for Estonian being included among mora-counting languages (Eek, Meister 2004:336), our system of features will be based on stress and quantity degree, considering the main characteristics of syllable and foot structure in Estonian language.

The central unit of Estonian prosody is the foot, carrying three quantity degrees (Q1, Q2, Q3) as the phonologically relevant prosodic oppositions. The quantity degree is a suprasegmental feature resulting from the joint effect of several other features, one of the most important of which is the ratio of the durations of syllables or their components[1]. Estonian quantity degrees and stress are usually described in the framework of a prosodic hierarchy enabling to divide an utterance into components lying on different levels of subordination (Eek, Meister 2004:253). As can be seen in Fig. 1, a sentence or phrase consists of prosodic words, while the words, in turn, consist of feet, the feet consist of syllables and the rear is brought up by phonemes, making up the lowest segmental level. As in Estonian sentences, phrases (noun phrase, verb phrase, adverbial phrase) are often quite closely intertwined, we define a phrase for the present paper as a finite clause or a list element bounded by an intra-sentence punctuation mark or conjunction. Thus, Fig. 1 may apply equally to either a sentence or a phrase.

The relative position of a current phone is rendered in a hierarchic scale as follows: position of the phone in the syllable, position of the syllable in the foot, position of the foot in the word, position of the word in the phrase. In addition, as has been proved by previous analysis, information on sentence and word length is necessary.

The next underlying principle of feature selection states that every phone has its intrinsic duration, while at the same time the phone is affected by its neighbouring phones. Intrinsic durations and phone interaction have been studied for many languages. Such universal linguistic phenomena are also observed in Estonian. The first measurements of the intrinsic durations of Estonian vowels took place about half

[1] The quantity degree is defined as a ratio of structural components of stressed and unstressed syllables in the form $\sigma_{stressed}$(nucleus+[coda]) / $\sigma_{unstressed}$(nucleus).

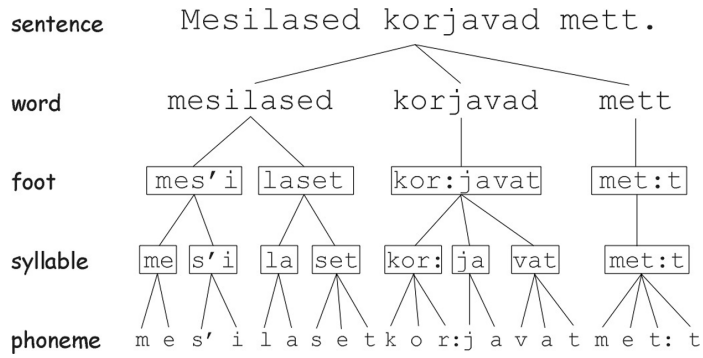


Figure 1. 'Bees gather honey' – hierarchical encoding of the relative position and length of a current speech unit. For example the phone [l] is encoded according to its position in the syllable [la] of a length of two phones. The position of the syllable [la] is encoded in relation to the foot [laset] with a length of two syllables. The foot is further assigned a code according to its place in the word [mesilased] etc.

a century ago (Liiv 1961). In several later studies of intra-phone microprosodic variations it has been stated that in Estonian the short low vowels are about 10–15 ms longer than the high vowels (Eek, Meister 2003:836; Meister, Werner 2006:111). Interaction between neighbouring phones is manifested in consonant shortening in clusters, in particular in the neighbourhood of voiceless consonants (Eek, Meister 2004:267).

The current phoneme segment is characterized by phoneme identity and phoneme length. In Estonian there are 9 vowel phonemes and 17 consonant phonemes (Eek, Meister 1999). The class and length of the left and right neighbours are also important. In modelling, the basic question is how many left and right neighbours affect the duration of a current phone. Usually their number is 1, 2 or 3. Experimental modelling of Estonian phone durations has shown that it takes two phonemes from the left (previous and previous but one) and two from the right (next and next but one) to achieve an optimal description of the context of a current phoneme (see Fig. 2). The phonemes are defined by their phoneme class (9 classes, pause included) and contrastive length (short vs. long). A phoneme and its context takes 10 features to describe, while the hierarchical position of the phoneme in the utterance is encoded by 5 features, some speech units (syllable stress, syllable type, foot quantity degree) take 3 features, and the length of higher-level units (syllable, foot, word, phrase, sentence) needs 5 features. There is also a binary feature referring to a punctuation mark. All these features (24 in all) make up a vector of basic features to serve as input for the durational model. Another point to consider when selecting initial features was their being supported by the technologies available for the Estonian language, enabling the features to be generated automatically from the input text. For the present study we have made use of a sentence builder, syllabifier, morphological analyzer, disambiguator a.o. modules provided by Estonian language technologists (Viks 2000), (Kaalep, Vaino 2001).

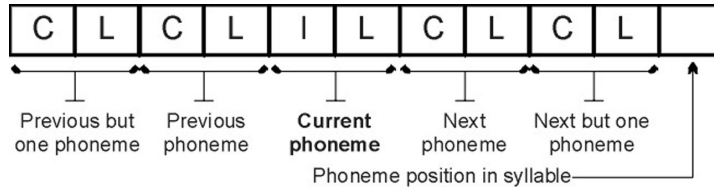


Figure 2. Encoding information on current phoneme neighbourhood (C – phoneme class, L – phoneme contrastive length, I – phoneme identity).

3.1. Initial data and the modelling environment

As our aim was to model speech temporal structure for a text-to-speech synthesizer, the analysis was based on read-out texts. A one-to-one correspondence between text and speech enables transition from a symbolic representation of prosody to an acoustic one as well as to find out to what extent, if at all, the syntactic structure of the written text could be related to the prosodic structure of speech.

The training material consisted of speech passages from a CD-version of a mystery story (Stout 2003) read by a professional actor, speech passages from longer news texts read by announcers of the Estonian Radio, and speech passages from the BABEL Estonian phonetic database (Eek, Meister 1999). In total there were over 60 speech samples read by 27 speakers , while the samples lasted from half to two minutes. All those samples were manually segmented into phones and pauses.

The speech temporal structure was modelled statistically using the Enterprise Miner workspace of SAS 9.1 software (see Fig. 3 for a block diagram of the data processing).

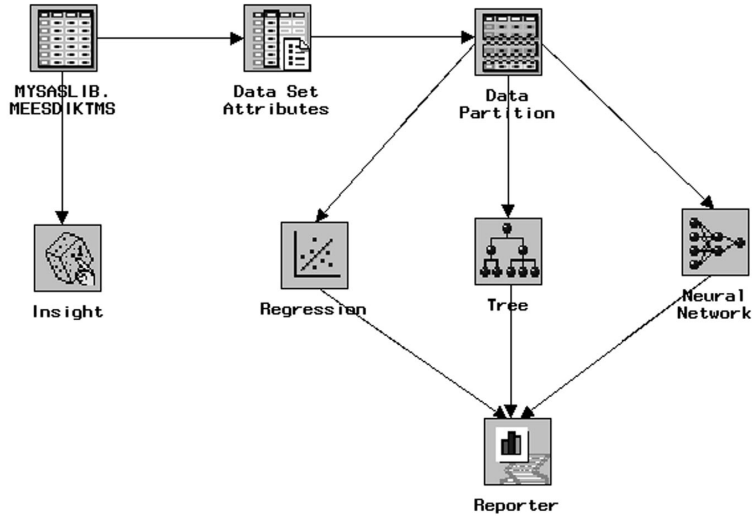


Figure 3. SAS Enterprise Miner workspace.

Linear regression, CART and neural networks were used as methods of prediction. In modelling, the material of each speaker's total presentation was divided into three parts: 50% was assigned for training the model, while 30% was used for data validation and the remaining 20% was meant for testing.

3.2. Expert opinions

After a rough selection has been made it would be quite useful to have a few experts give their opinions on the resulting vector of argument features as well as a few recommendations on possible additional features. The experts could estimate whether or not a feature is significant in the prediction of speech timing (e.g. segmental durations), also, it would be useful to have their opinion on the possible joint effects of certain feature constellations. During our first experiments in statistical modelling we invited Estonian phoneticians and speech technologists to evaluate our first vector of argument features. The coincidence between their opinions and our preliminary results was a mere 41–65% (Mihkla, Kuusik 2005:95). Due to the interim accumulation of speech material, however, and the increasing volume of Estonian prosodic speech corpora our recent results are in better harmony with expert opinions (see Table 1). Part of the – still considerable – difference may be due to the fact that most of the 'duration patterns' of the phoneticians are based on measurements of isolated words and sentences, whereas our results draw on fluent speech. Sound

Table 1. Expert opinions versus results of regression analysis (ExpN- N expert, Reg – results of regression analysis, 1 – significant explanatory variable, 0 – insignificant variable)

Explanatory variable	Exp1	Exp2	Exp3	Exp4	Reg
Previous phoneme class	0	0	0	0	1
Previous phoneme length	1	1	1	1	1
Current phoneme identity	1	1	1	0	1
Current phoneme length	1	1	1	1	1
Next phoneme class	1	1	0	0	1
Next phoneme length	1	0	1	1	0
Phoneme position in syllable	1	1	0	1	1
Stress of syllable	1	1	1	1	1
Type of syllable	1	0	0	1	0
Quantity degree of foot	1	1	1	1	1
Syllable position in foot	1	1	1	1	1
Length of foot in syllables	1	1	0	1	1
Foot position in word	1	0	0	1	0
Length of word in feet	1	1	0	1	1
Word position in phrase	1	1	1	1	1
Length of phrase in words	1	0	0	1	1
Length of sentence in phrases	1	0	0	0	1
Total of 'correct' answers	13	14	9	10	
%	76%	82%	53%	59%	
Total average %				67%	

durations measured on isolated sentences and the temporal structure of fluent speech, however, are known to differ quite considerably (Campbell 2000:312–315).

3.3. Lexical prosody

Traditionally a list of factors significantly affecting speech timing does not include either part-of-speech (POS) information or morphological characteristics (van Santen 1998, Campbell 200, Sagisaka 2003). This may be due to most studies on TTS synthesis focusing on languages with relatively little morphology. Finnish is one of the few languages boasting a study of the influence of morphological features on the duration of speech units (Vainio 2001). In Estonian the word has a very important role both in grammar and phonetics, while the morphology is extremely rich. Hence our interest to check whether there are any morphological, lexical, or even syntactic features possibly affecting the temporal structure of Estonian speech. The most natural way to find out that information seemed to lie through an extension of our earlier methodology of statistical modelling to see how certain morphological, lexical and syntactic characteristics might affect the functioning of our durational models. The modelling was done using two different methods - linear regression and a nonlinear method of neural networks. Change of the output error was measured to enable qualitative assessment of the influence of the factors under study. The results demonstrated a couple of percent error decrease in case some morpho-syntactic and POS information had been added to model input (Mihkla 2007).

As the models were based on the speech of merely two radio announcers it seemed a little premature to generalize the possible interpretations. It should however be mentioned that the most distinct regularities were revealed by a visual observation of the POS regression coefficients. Table 2 demonstrates the mean lengthenings and shortenings, by part of speech, of speech sounds relative to verb sounds in the durational models of male and female speech. As we can see, there is more variation in the middle part of the table, while the top and bottom parts are rather similar. Table 2 reveals that in proper names sounds are pronounced longer by 5.2–6.2 ms on average. The two newscasters' mean phone length was 62.5 and 64.1 ms respectively. Consequently their pronunciation of proper names was about 10% longer than verbs. Of the latter, nouns and adpositions were pronounced a little longer. It was surprising to find such lengthening in adpositions as in most languages function words are shorter than content words. An Estonian adposition invariably belongs to a noun phrase. The noun often stands in the focus of the sentence, while its more than average length may extend to a neighbouring adposition. Ordinal numerals, however, were pronounced over 10% shorter, and pronouns and adverbials ca. 5% shorter than average. The shortening of ordinal numerals can be accounted for by quite many dates in the text, which are typically expressed by ordinal numerals. Reading the relatively long dates of the past century the newscasters tend to hurry. This is because usually only the last one or two numbers of the year are important, but nevertheless the whole number has to be pronounced, as required by rules of correct reading.

Table 2. The average lengthening-shortening values (in ms) of sound durations for different parts of speech in the male and female material

Part of speech	Male announcer	Female announcer
Proper noun	6.23	5.22
Noun	2.25	2.10
Adposition	0.82	2.82
Genitive attribute	0.42	1.35
Verb	0.00	0.00
Numeral	–0.10	0.42
Conjunction	–0.14	1.81
Adjective	–0.39	1.14
Adverb	–0.89	–2.90
Pronoun	–4.13	–3.86
Ordinal numeral	–5.44	–7.48

3.4.Pauses in speech

Regulating the primary syntactic division or prosodic phrasing with an aim to facilitate comprehension of the utterance, pauses are one of the factors determining speech rhythm (Tseng 2002). Despite the high variability of pauses in natural speech, different kinds of pauses vary in duration. At least in texts read aloud at a normal speech rate it is possible to distinguish between phrase-, sentence- and paragraph-final pauses by their duration, as has been proved by statistical analysis (Mihkla 2006a:290). A natural-sounding rhythm of synthetic speech would mean a good enough rendering of both the duration of pauses and their location in the speech flow.

For modelling pause durations some texts were used to generate a number of characteristics describing the following:

- text structure (end of paragraph, sentence, or phrase; conjunctions within the text)
- prepausal foot (length of the foot in phones, foot quantity degree, length of the foot-final syllable in phones and a binary characteristic indicating final lengthening)
- pause timing specifications (distance of the pause from the beginning of the paragraph, sentence, and phrase, as well as its distance from the previous pause and the previous breathing).

The feature to be predicted was pause duration. The use of linear regression required the response to be logarithmed, for logarithmed values are more likely to yield a normal distribution. See Fig. 4 for a regression tree to calculate pause duration.

The first-level classification of pauses on the regression tree distinguishes between sentence-final and non-sentence-final pauses. The intra-sentence pauses branch off to the left, while sentence-final ones go to the right. The intra-sentence pauses are, in turn, dichotomized depending on whether they happen to finish a phrase or not. The length of an intra-phrasal pause, for example, distanced from the previous pause by less than seven feet is 166 ms, whereas a longer distance correlates with a 253 ms pause.

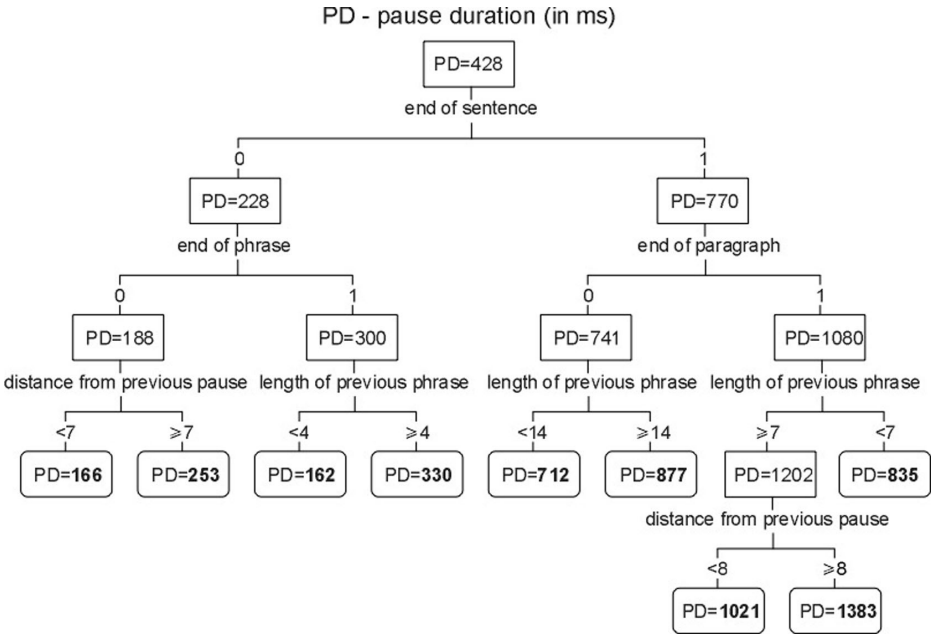


Figure 4. Regression tree to model pause durations in a text. The boxes with rounded corners stand for the leaves of the tree, each referring to the duration of the appropriate pause class.

Prediction of pause location was started by applying logistic regression, meant to predict the probability of a pause following a given word in the speech flow. The input variables were the same as used for predicting pause durations. There were two additional binary features, indicating whether the following word is a proper noun or a foreign word. Their addition was inspired by the idea that there might be a short pause before the pronunciation of proper names (e.g. *Minu nimi on Tamm, Jüri Tamm* 'My name is Tamm, Jüri Tamm') and maybe also before some more sophisticated foreign words (e.g. *Rahvas toetas konstitutsioonilist monarhiat* 'The people supported constitutional monarchy'). That hypothesis was, however, disproved. The correlation of pauses with proper names and foreign words was extremely weak and thus the features proved insignificant.

Table 3. Pause locations as predicted by the logistic model

	Correct predictions	Actual no. of pauses in the sample	Percent of correctness
Pause after word (PAUSE = 1)	402	600	67
No pause (PAUSE=0)	2510	2708	93
TOTAL			88

As can be seen from Table 3 the model predicted correctly the location of 67% of the pauses. The total predictive precision of the model, i.e. its ability to guess whether a certain word will be followed by a pause or not, amounted to 88%. Analysis showed that certain markers of text structure (end of paragraph, end of sentence, colon, dash) are followed by a pause with a 93-100% probability. Leaving out such well-marked pauses the model was left with a mere 44% of predictive power.

3.5. Feature significance, predictive precision and errors

Prior to evaluating the significance of features one should test the statistical relevance of the model. Table 4 shows an example of the summary of fit and variability analysis of our regression model of pause durations. We can see that the model is statistically significant and it describes ca 67% of the variability of pause durations (R-square = 0.6686). Other durational models described 65–73% of pause duration variability, while for phone durations the reading amounted to 52–63%.

Tabel 4. Summary of fit and variability analysis of a regression model of pause durations

Summary of fit					
Mean of response –0,86673			R-square 0.6686		
Analysis of variance					
Source	DF	Sum of squares	Mean square	F stat	Pr > F
Model	9	478.5	18.403	220.87	<0.0001
Error	560	408.8	0.0862		
C Total	559	787.2			

Feature significance is best determined in the case of a regression model, where the significance of each feature gets a statistical evaluation. The method of forward selection, for example, means that the most significant features at the moment are added to the model one by one, with revaluation taking place before each cycle. CART also makes sure that the regression tree will get the most significant features. In neural networks, however, there is no such evaluation of significance and the relevance of a feature can be estimated manually, by adding or removing features one by one and evaluating the output for each case.

Table 5 contains the features that – on the basis of extensive experimental material – have been found to be significant in predicting phone duration. The number of significant features may vary across speakers; it also depends on the used method. The features in bold letters on dark grey background mark features that were significant for all speakers. The features in normal lettering on a lighter grey background were insignificant for the durational models of some speakers. Surprisingly the latter group includes the quantity degree of foot, which is considered a cornerstone of Estonian speech prosody. One of the possible reasons might lie in the

circumstance that quantity degree as a suprasegmental feature cannot be represented as a single linear characteristic, but rather as a ratio of structural components of

Table 5. Significance of input features for modelling segmental durations

	Inputs (per *sound)*
1.	**previous but one phoneme class**
2.	**previous but one phoneme length**
3.	**previous phoneme class**
4.	previous phoneme length
5.	**current phoneme identity**
6.	**current phoneme length**
7.	**next phoneme class**
8.	next phoneme length
9.	**next but one phoneme class**
10.	**next but one phoneme length**
11.	**phoneme position in syllable**
12.	stress of syllable
13.	type of syllable
14.	quantity degree of foot
15.	**syllable position in foot**
16.	**length of foot in syllables**
17.	foot position in word
18.	**length of word in feet**
19.	**monosyllabic word**
20.	**word position in phrase**
21.	**length of phrase in words**
22.	length of sentence in phrases
23.	**punctuation**
24.	**morphology**
25.	**part-of-speech**
26.	**part of sentence**

stressed and unstressed syllables in the form $\sigma_{stressed}$(nucleus+[coda]) / $\sigma_{unstressed}$(nucleus). Another fact suggestive of possible mutual effects between stress and syllable structure is that syllable stress does not always correlate significantly with the duration to be predicted. Surprisingly enough the contrastive length of the next phone happened to be less significant than that of the next but one. The normal type on a white background marks the two features that proved systematically insignificant for the prediction of segmental durations, notably, type of syllable (open or closed) and the position of the foot in the word.

Table 6 contains the six features proved by logistic regression to affect the positioning of an intra-sentence pause. A comma in the text is very important, raising the chances for a pause to occur in speech by 17.4 times. A 7–8 times higher chance for the word to be followed by a pause is signalled by a following conjunction or a lengthened final foot. Slightly more frequently than average a pause can be expected to occur after longer feet or after words of longer quantity degrees. In a predictive model, however, the role of the latter two features remains relatively marginal, raising the chances for a pause to occur by no more than 1.2–1.3 times.

Table 6. Results of logistic regression: variables explanatory of the location of an intra-sentence pause, their ratio of chances and confidence levels

Independent variables	Odds ratio	Confidence levels	
		Lower	Upper
The word is followed by a comma	17.4	11.7	25.9
The next word is a conjunction	7.9	4.8	12.8
Distance of the word from sentence beginning	1.1	1.0	1.2
Length of the preceding foot	1.3	1.1	1.4
Quantity degree of the preceding foot	1.2	1.1	1.5
Lengthening of the preceding foot	6.9	5.2	9.2

Depending on the speaker and the method, the predictive error of phone durations was within 16.1–21.2%. Testing different methods (linear regression, CART, neural networks) on one and the same data set, it turned out that the predictive precision of linear regression and the neural networks was nearly equal, while the CART model had a slightly higher predictive error than the rest (Mihkla 2006b:123). We were surprised to find that the linear method could compete with non-linear ones, although a linear model is usually expected to show nothing but the most obvious and most general relations between input and output, and only nonlinear methods are trusted to reveal the more covert relations.

For pause durations the predictive error was - depending on the speaker - 8-12%. The predictive precision of the logistic model for pause locations is presented in Table 3. If, however, the punctuation-bound pauses are left out, the predictive precision of intra-sentence pauses drops to 44%.

4. Conclusion

Generation of synthetic speech with a prosodically appropriate temporal structure is complicated as speech prosody is influenced by many factors. In feature selection one should consider certain general aspects governing speech timing as well as some language specific ones. The Estonian durational model stands out for certain foot-bound features (foot quantity degree, number of feet in the word) being included in the model input. Although feature selection involved the use of expert opinions, too, the vector of optimal argument features should definitely be tested by statistical methods. The considerable difference between the expert opinions and the results of fluent speech analysis may be due to the fact that the 'durational patterns' underlying the decisions of the expert phoneticians were quite likely based on measurements of laboratory speech (i.e. isolated sentences and words). Prediction of pause durations and locations in the speech flow was proved to heavily depend on the following features: distance of the word from sentence beginning and the previous pause, length

of the preceding phrase, length and quantity degree of the preceding foot, and the occurrence of a punctuation mark or conjunction. For the durations of speech units the predictive precision of the model differed across the methods used as well as speakers, while for segmental durations the predictive error was 16–21% and for pause durations it was 8–12%. Although in the logistic model the summary estimation of the possible occurrence of a pause after a text word amounted to 88%, predictive precision for intra-sentence pauses did not exceed 44%.

Apart from the traditional parameters, describing the context of a phone and its hierarchic position in the sentence, prediction of Estonian segmental durations requires an addition of certain morphological, syntactic and lexical features such as word form, part of sentence and part of speech. The most distinct regularities were revealed by part-of-speech analysis of the durational model, showing that proper names and nouns were pronounced the longest, whereas the least time was spent on ordinal numerals and pronouns.

Acknowledgements

The study was financed by the national programme „Language technological support of Estonian". My heartfelt thanks to Arvo Eek (Institute of Cybernetics at TTU, Tallinn) and Einar Meister (Institute of Cybernetics at TTU, Tallinn) for their valuable remarks and comments.

Address:
Meelis Mihkla
Institute of the Estonian Language
Roosikrantsi 6
10119 Tallinn, Estonia
Tel.: +372 6446 947
E-mail: meelis@eki.ee

References

Campbell, Nick (2000) "Timing in speech: a multilevel process". In *Prosody: theory and experiment*, 281–334. M. Horne, ed. Dordrecht/Boston/London: Kluwer Academic Publishers.

Campbell, N. W. and S. D. Isard (1991) "Segment durations in a syllable frame" *Journal of Phonetics* 19, 37–47.

Eek, Arvo and Einar Meister (1999) "Estonian speech in the BABEL multi-language database: phonetic-phonological problems revealed in the text corpus". In *Proceedings of LP'98*, II, 529–546. O. Fujimura, ed. Prague: The Karolinum Press.

Eek, Arvo and Einar Meister (2003) "Foneetilisi katseid ja arutlusi kvantiteedi alalt (I): Häälikukestusi muutvad kontekstid ja välde". [Phonetic tests and disputes about quantity (I): Contexts changing sound duration and quantity degree.] *Keel ja Kirjandus* (Tallinn) 46, 11, 815–837 and 12, 904–918.

Eek, Arvo and Einar Meister (2004) “Foneetilisi katseid ja arutlusi kvantiteedi alalt (II): Takt, silp ja välde”. [Phonetic tests and disputes about quantity (II). Foot, syllable and quantity.] *Keel ja Kirjandus* (Tallinn) 47, 4, 251–277 and 5, 336–357.

Dutoit, Thierry (1997) *An introduction to text-to-speech synthesis*. Dordrecht: Kluwer Academic Publishers.

Horak, Pavel (2005) “Using neural networks to model Czech text-to-speech synthesis”. In *Proceedings of the 16th Conference of electronic speech signal processing*, 76–83. R. Vich, ed. Prague: TUDpress.

Huggins, A.W.F. (1968) “The perception of timing in natural speech: compensation within syllable”. *Language and Speech* 11, 1–11.

Kaalep, Heiki-Jaan and Tarmo Vaino (2001) “Complete morphological analysis in the linguist’s toolbox”. In *Congressus Nonus Internationalis Fenno-Ugristarum, Tartu 7.-13.08.2000*, V, 9–16. Tartu: TÜ Kirjastus.

Klatt, D. H. (1979) “Synthesis by rule of segmental durations in English sentences”. In *Frontiers of Speech Communication research*, 287–300. B. Lindblom and S. Öhman, eds. New York: Academic Press.

Liiv, Georg (1961) “Eesti keele kolme vältusastme vokaalide kestus ja meloodiatüübid”. [Duration of vowels of the three quantity degree of Estonian and types of melody.] *Keel ja Kirjandus* (Tallinn) 4, 7, 412–424 and 8, 480–490.

Meister, Einar and Stefan Werner (2006) “Intrinsic microprosodic variations in Estonian and Finnish: acoustic analysis”. In *Fonetiikan Päivät 2006 = The Phonetics Symposium 2006*, 103–112. R. Aulanko, L. Wahlberg, and M. Vainio, eds. (Publications of the Department of Speech Sciences, University of Helsinki) Helsinki: University of Helsinki.

Mihkla, Meelis and Jüri Kuusik (2005) “Analysis and modelling of temporal characteristics of speech for Estonian text-to-speech synthesis”. *Linguistica Uralica* 41, 2, 91–97.

Mihkla, Meelis (2006a) “Pausid kõnes”. [Pauses in Speech.] *Keel ja Kirjandus* (Tallinn) 49, 4, 286–295.

Mihkla, Meelis (2006b) “Comparison of statistical methods used to predict segmental durations”. In Fonetiikan Päivät 2006 = The Phonetics Symposium 2006, 120–124. R. Aulanko, L. Wahlberg, and M. Vainio, eds. (Publications of the Department of Speech Sciences, University of Helsinki) Helsinki: University of Helsinki.

Mihkla, Meelis (2007) “Morphological and synthetic factors in predicting segmental durations for Estonian text-to-speech synthesis”. *Proceedings ICPhS 2007*. (accepted, in print).

Sagisaka, Yoshinori (2003) “Modeling and perception of temporal characteristics in speech”. In *Proceedings of 15th International Congress of Phonetic Sciences*, 1–6. M. J. Sole, D. Recasens, and J. Romero, eds. Barcelona.

van Santen, Jan (1998) “Timing”. In *Multilingual text-to-speech synthesis: the Bell Labs approach*, 115–140. R. Sproat, ed. Kluwer Academic Publishers.

Stout, Rex 2003 “Deemoni surm”. [Death of a Demon.] CD-versioon (Read by Andres Ots). Tallinn: Elmatar.

Tatham, Mark and Katherine Morton (2005) *Developments in speech synthesis*. Chichester: John Wiley & Sons Ltd.

Tseng, C. (2002) “The prosodic status of breaks in running speech: examination and evaluation”. In *Proceedings of Speech Prosody 2002*, 667–670. Aix-en-Provence, France.

Vainio, Martti (2001) *Artificial neural network based prosody models for Finnish text-to-speech synthesis*. Helsinki: University of Helsinki.

Viks, Ülle (2000). “Eesti keele avatud morfoloogiamudel” [Open morphology model of Estonian language.]. In *Arvutuslingvistikalt inimesele*, 9–36. [From computational linguistics to people.] T. Hennoste, ed. (Tartu Ülikooli üldkeeleteaduse õppetooli toimetised, 1.) Tartu.

CURRICULUM VITAE

Meelis Mihkla

Sündinud 6. juulil 1955 Tallinnas
Kodakondsus: Eesti
Abielus, neli tütart
Aadress: Eesti Keele Instituut
Roosikrantsi 6, 10119 Tallinn
Telefon: +372 6177544
E-post: Meelis.Mihkla@eki.ee

Haridus

1973–1978	Tallinna Tehnikaülikool, automatiseeritud juhtimissüsteemide eriala
2001	Tartu Ülikool, MA eesti keele erialal
2005–2007	Tartu Ülikool, doktorikool „Keeleteadus ja –tehnoloogia”

Teenistuskäik

1977–1980	Keele ja Kirjanduse Instituut, insener
1980–1992	Küberneetika Instituudi Arvutustehnika Erikonstrueerimisbüroo, insener-programmeerija
1993–2006	Eesti Keele Instituut, haldusdirektor
2007–…	Eesti Keele Instituut, teadur

Teadustegevus

Uurimisvaldkonnad: kõneprosoodia, kõneüksuste andmebaasid
Publikatsioone: ca 30
Uurimistoetused:

1996–97	ETF grant nr 1995 „Eesti keele tekst-kõne süntesaator“, grantihoidja
1998–99	ETF grant nr 3511 „Grafeem-foneem teisendus ja prosoodia modelleerimine eesti keele kõnesüntesaatorile“, grantihoidja
1998–99	AEF stipendiaat „Kõnesünteesi liidesed inimestele arvutiga suhtlemiseks ja Internetist informatsiooni saamiseks“
2000–01	ETF grant nr 4194 „Kõnesünteesi kvaliteedi hindamine ja kasutajaliidesed eestikeeelsele tekst-kõne süsteemile“, grantihoidja

2001–02 EL Phare Access projekt „Eesti tekst-kõne süntesaator pimedatele“, projektijuht

2002–05 ETF grant nr 6912 „Süntaktiliste ja prosoodiliste tunnuste ühilduvus kõnesünteesis“, põhitäitja

Teaduslik organisatsiooniline ja erialane tegevus

1996–2006 Eesti Keele Instituudi teadusnõukogu liige

2004–… Riikliku programmi „Humanitaar- ja loodusteaduslikud kogud” juhtkomitee liige

2006–… Emakeele Seltsi liige

2007–… Eesti Rakenduslingvistika Ühingu liige

CURRICULUM VITAE

Meelis Mihkla

Born on July 6, 1955 in Tallinn
Citizenship: Estonian
Married, four daughters
Addres: Institute of the Estonian Language
Roosikrantsi 6, 10119 Tallinn
Telephone: +372 6177544
E-mail: Meelis.Mihkla@eki.ee

Education

1973–1978	Tallinn University of Technology, automatic control systems
2001	University of Tartu, MA Estonian language
2005–2007	University of Tartu, doctoral school „Linguistics and language technology”

Professional experience

1977–1980	Institute of Language and Literature, engineer
1980–1992	Special Design Bureau of Computational Technology at the Institute of Cybernetics, engineer-programmer
1993–2006	Institute of the Estonian Language, assistant director
2007–	Institute of the Estonian Language, researcher

Research interests and grants

Research areas: Speech prosody, databases of speech units
Number of publications: ca 30
Grants/projects:

1996–97	ESF grant “Estonian text-to-speech synthesizer”, principal investigator
1998–99	ESF grant “Grapheme-phoneme transcription and prosody modeling for the Estonian speech synthesizer”, principal investigator
1998–99	OEF project “Interfaces of speech synthesis for blind people”, principal investigator

2000–01 ESF grant “Evaluation the quality of speech synthesis and interfaces for Estonian TTS”, principal investigator
2001–02 Phare Access project “Estonian Text-to-Speech Synthesizer for the Blind”, co-investigator
2002–05 ESF grant “Congruence of syntactic and prosodic features in speech synthesis”, co-investigator

Research-administrative experience

1996–2006 Member of Scientific Council of the Institute of The Estonian Language
2004– Member of leading committee of state program “Humanitarian and natural science archives”
2006– Member of Mother Tongue Society
2007– Member of Estonian Applied Linguistics Society

DISSERTATIONES LINGUISTICAE UNIVERSITATIS TARTUENSIS

1. **Anna Verschik.** Estonian yiddish and its contacts with coterritorial languages. Tartu, 2000, 196 p.
2. **Silvi Tenjes.** Nonverbal means as regulators in communication: socio-cultural perspectives. Tartu, 2001, 214 p.
3. **Ilona Tragel.** Eesti keele tuumverbid. Tartu, 2003, 196 p.
4. **Einar Meister.** Promoting Estonian speech technology: from resources to prototypes. Tartu, 2003, 217 p.
5. **Ene Vainik.** Lexical knowledge of emotions: the structure, variability and semantics of the Estonian emotion vocabulary. Tartu, 2004, 166 p.
6. **Heili Orav.** Isiksuseomaduste sõnavara semantika eesti keeles. Tartu, 2006, 175 p.
7. **Larissa Degel.** Intellektuaalsfäär intellektuaalseid võimeid tähistavate sõnade kasutuse põhjal eesti ja vene keeles. Tartu, 2007, 225 p.

DISSERTATIONES LINGUISTICAE UNIVERSITATIS TARTUENSIS
8

KÕNE AJALISE STRUKTUURI MODELLEERIMINE EESTIKEELSELE TEKST-KÕNE SÜNTEESILE

MODELLING THE TEMPORAL STUCTURE OF SPEECH FOR THE ESTONIAN TEXT-TO-SPEECH SYNTHESIS

MEELIS MIHKLA

TARTU 2007

MEELIS MIHKLA

KÕNE AJALISE STRUKTUURI MODELLEERIMINE EESTIKEELSELE TEKST-KÕNE SÜNTEESILE

ISSN 1024–395X
ISBN 978–9949–11–797–0