T A R T U Ü L I K O O L

MATEMAATIKA-INFORMAATIKATEADUSKOND

Arvutiteaduse instituut

Informaatika eriala

Tõnis Nurk

Markovi peitmudelitel põhineva häälemudeli loomine eestikeelse kõnesünteesi jaoks

Bakalaureusetöö (6 EAP)

Juhendaja: Meelis Mihkla (Eesti Keele Instituut)

Autor: …… "….." mai 2012

Juhendaja: ……. "….." mai 2012

Lubada kaitsmisele

Professor: ……. "….." mai 2012

TARTU 2012

Sisukord

Sissejuhatus

Kirjaliku teksti automaatset teisendust suuliseks kõneks nimetatakse tekst-kõne sünteesiks ehk lihtsustatult kõnesünteesiks. Alates eelmise sajandi lõpust on kõnesünteesisüsteeme arendatud paljude inimese ja masina vahelise suhtlemise liideste väljundseadmeteks. Esimesed rakendused olid enamjaolt mõeldud puuetega inimeste elu hõlbustamiseks, enamlevinud kasutusel olevaks tooteks peetakse peamiselt nägemispuuetega inimestele mõeldud ekraanilugejat. Viimasel kümnendil on kõnesüntees kanda kinnitanud ka laiemal turul, seda peamiselt tänu sünteesikvaliteedi tõusule ja arvutusressursside kättesaadavusele, näiteks automaatselt genereeritud audioraamatud, kõneposti ja e-kirjade ettelugemine, kombineerituna kõnetuvastusega liidesed automaatseks infootsinguks telefoni kaudu, dialoogsüsteemid jne. Eestis arendatakse kõnesünteesi riiklike programmide raames Eesti Keele Instituudis [1] ja Tallinna Tehnikaülikooli Küberneetika Instituudis [2].

Kõnesünteesitehnoloogiaid on kasutusel mitmeid. Konkatenatsioonil põhinevad tehnoloogiad hõlmavad endas korpusest üksuste valiku algoritmidega kõne genereerimise, difoonsünteesi[1] ja domeenipõhist sünteesi, kus sünteesil liidetakse kokku eelnevalt salvestatud sõnad või fraasid (nt rääkivad kellad, kalkulaatorid, autode navigatsioonisüsteemid). Formantsünteesis püütakse modelleerida häälekurdude tööd ning kõnetrakti resonantssagedusi ehk formante. Artikulatoorse kõnesünteesi puhul tekitatakse kõne digitaalselt, simuleerides õhuvoo liikumist läbi kõnetrakti esituse. Statistilised parameetrilised mudelid, millele antud töös keskendutakse, baseeruvad Markovi peitmudelitel või sellega lähedastel mudelitel ja võeti kõnesünteesi jaoks kasutusele 1990ndate keskel Jaapanis [3].

Töö põhieesmärgiks on luua Markovi peitmudelite abil treenitud häälemudelid[2] mees- ja naishäälele, mida on võimalik kasutada eestikeelse kõnesünteesi rakendustes. Eesmärgi saavutamiseks antakse töös esmalt ülevaade vajalikest mõistetest ja ressurssidest. Häälemudeli treenimine realiseeritakse süsteemiga HTS (HMM-based Speech Synthesis System) [4] ja kohandatakse eestikeelse kõnesünteesi moodulitega, mis on arendatud/realiseeritud keskkonnas Festival (The Festival Speech Synthesis System) [5]. Häälemudelite treeningkorpused koostatakse empiiriliselt hinnatud valikkriteeriumide järgi.

[1] Difoon – foneemipaar, nt isoleeritud sõnas „oja“ esinevad difoonid „#o“, „oj“, „ja“ ning „a#“, kus # tähistab vaikust või pausi.

[2] Kui räägitakse kõnesünteesiprotsessis kasutatavast kõne loomise mudelist üldiselt või selle omadustest, kasutatakse mõistet kõnemudel. Konkreetset treenitud või planeeritavat mudelit nimetatakse häälemudeliks.

Töö koosneb järgmistest osadest. Esimeses peatükis antakse kirjanduse põhjal ülevaade kõnesünteesisüsteemi toimimisest ja tuuakse näited üldkasutatavamatest paradigmadest. Teises peatükis kirjeldatakse täpsemalt statistilist parameetrilist kõnesünteesi ning seletatakse lahti käesolevas töös kasutatud sünteesisüsteemi põhimõtted. Kolmandas peatükis käsitletakse Eesti Keele Instituudi näitel kõnesünteesi jaoks kõnekorpuse loomise protsessi. Neljas peatükk on pühendatud kõnemudeli loomisele, kirjeldatud on süsteemi HTS kohandamine eesti keelele, treenitud häälemudelite ja seda mõjutavate tegurite hindamine. Välja on toodud järgnevad võimalikud tegevused kasutatud sünteesimeetodil genereeritud kõne kvaliteedi parandamiseks.

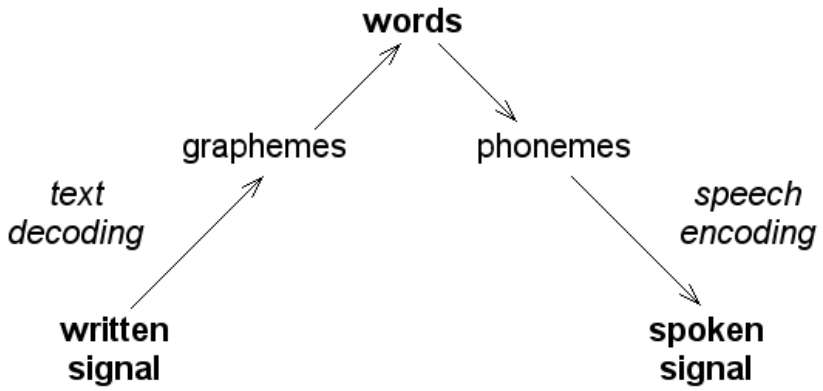
Lisades 1 ja 2 on toodud süsteemis HTS kasutusel olev foneetiline tähestik ja foneemiklassifikatsioon häälemudeli „Liisi“ näitel. Lisas 3 on CD-plaat erinevate häälemudelitega genereeritud näitelausetega.

Lugejalt eeldatakse lisaks arvutiteaduslikule taustale keeleteaduse ja arvutuslingvistika põhimõistete tundmist.

Kõnekorpus ja Markovi peitmudelitel põhinevad häälemudelid on loodud Eesti Keele Instituudis projektide „Eestikeelne korpuspõhine kõnesüntees“ ja „Kõnesünteesiliidesed“ raames. Autoril on käesolevas töös kajastatud mudelite loomisel olnud kandev roll, kõnekorpuse loomine ja hindamine on toimunud projekti „Eestikeelne korpuspõhine kõnesüntees“ põhitäitjate Meelis Mihkla, Indrek Kiisseli, Tõnis Nurga ja Liisi Piitsi ühise tööna.

1. Kõnesüntees

Tekst-kõne süntees on inimlugemise analoog. Süntesaatori sisendiks on tekst ja väljundiks genereeritud helisignaal. Kõnesünteesisüsteemis on üldistatuna kaks komponenti: tekstianalüüs ja signaalitöötlus [6]. Eeskomponent (*front end*) teisendab sisendteksti nö lingvistilisse spetsifikatsiooni, dekodeerides kirja pandud teksti häälduspärasesse vormi, ja teine komponent genereerib selle info põhjal heliväljundi, nagu on näha joonisel 1.1. Selline jaotus on praktiline nii teoreetiliselt kui praktilise teostuse poolest: eeskomponent, mis sisaldab tekstianalüüsi, on tüüpiliselt keelespetsiifiline, samas helisignaali töötluse moodulid on suuresti keelest sõltumatud (välja arvatud andmed, mida nad sisaldavad või millel treenitud on) [7].



Joonis 1.1: Tekst-kõne sünteesisüsteemi mudel [6].

1.1. Lingvistiline töötlus

Iga uudse lause sünteesil ennustab eeskomponent teksti põhjal lingvistilise spetsifikatsiooni. Lingvistiline info, mis antakse sisendiks sünteesiprotseduurile, võib sisaldada vaid foneemijärjendit, kuid parema tulemuse saamiseks peab lisama segmendiülest informatsiooni, nagu näiteks loodava heli prosoodiamuster. Teisisõnu, lingvistiline spetsifikatsioon sisaldab endas tegureid, mis võivad lauset moodustavate genereeritud helide akustilist realisatsiooni mõjutada [7].

Võtame näiteks vokaali „e“ sõnas „mees“. Lingvistiline spetsifikatsioon peab hõlmama kogu informatsiooni, mis võib mõjutada antud vokaali kõla, ehk on kokkuvõte antud konteksti kogu info kohta, milles vokaal esineb. Praegusel juhul kuuluvad konteksti näiteks eelnev konsonant „m“ (kuna mõjutab vokaali formanditrajektoore) ja fakt, et tegemist on ühesilbilise sõnaga (mõjutab vokaali kestust ja väldet). Lisaks võivad kõla mõjutada

muuhulgas tabelis 1.1 toodud tegurid. Nimekirja võib mõnel juhul lisada ka paralingvistilisi tegureid, näiteks kõneleja tuju või kuulaja isikupära. Praktilistel kaalutlustel piirdutakse siiski ühe lause piires esineva infoga.

Tabel 1.1: Lingvistilises spetsifikatsioonis sisalduda võivad kontekstitegurid.

Eelnevad ja järgnevad foneemid
Segmendi asukoht silbis
Silbi asukoht sõnas ja fraasis
Sõna asukoht fraasis
Eelneva/praeguse/järgneva silbi rõhk/pikkus/välde
Kaugus rõhulisest silbist
Eelneva/praeguse/järgneva sõna leksikaalne klass
Eelneva/praeguse/järgneva fraasi pikkus
Fraasi lõputoon
Lause pikkus mõõdetuna silpides/sõnades/fraasides

Tegurite loend on väga pikk ning arvestades erinevaid väärtusi, mida iga faktor saada võib, on erinevate võimalike kontekstide arv väga suur. Näiteks foneem võib eesti keele puhul olenevalt foneetikakirjeldusest omada kuni 100 väärtust. Lisas 1 toodud näites häälemudeli „Liisi“ puhul on kasutatud 59 foneemimärgendit[3], lisaks kaht tüüpi pause.

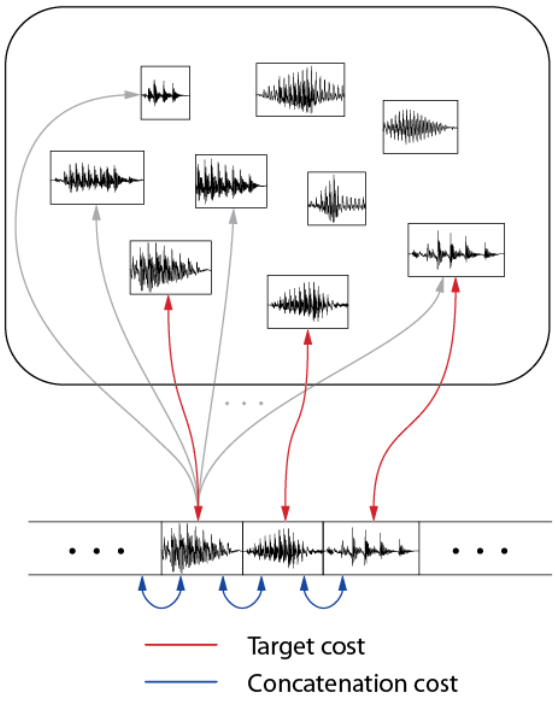
On teada, et kõik tegurid ei avalda mõju igal ajahetkel. Vastupidi, eeldatakse, et ühel ajahetkel avaldavad märkimisväärset mõju vaid üksikud tegurid. Antud teemat statistilise parameetrilise kõnesünteesi puhul käsitleme peatükis 3.2.

1.2. Näitepõhised süsteemid

Näitepõhine kõnesünteesisüsteem vajab sünteesiprotsessi toimimiseks kõnekorpust, mis on märgendatud foneetilises transkriptsioonis, et sellest saaks sünteesifaasis vajalikke segmente valida. Tüüpilise üksuste valikul põhineva sünteesi puhul sisaldab märgendus nii foneetilist kui prosoodilist infot. Segmentide valik sünteesil ei ole triviaalne, kuna korpuses ei ole garanteeritud kõikide spetsifikatsioonide esinemine, seega tehakse valik mitme mõningal määral ebasobiva segmendi hulgast, koostamaks konkateneerimiseks parim saadaolev järjend.

Joonisel 1.2 on kirjeldatud tüüpilise näitepõhise süsteemi, üksuste valikul põhineva sünteesi, helisignaalide sobitamise skeem, mille puhul püütakse minimeerida soovitule sarnaste segmentide valiku ja ühilduvuse kombineeritud maksumust.

[3] Kõnesünteesi jaoks on lisaks traditsioonilisele 26 foneemile [8] võetud kasutusele poolvokaal *w* ja ninahäälik *ŋ*.



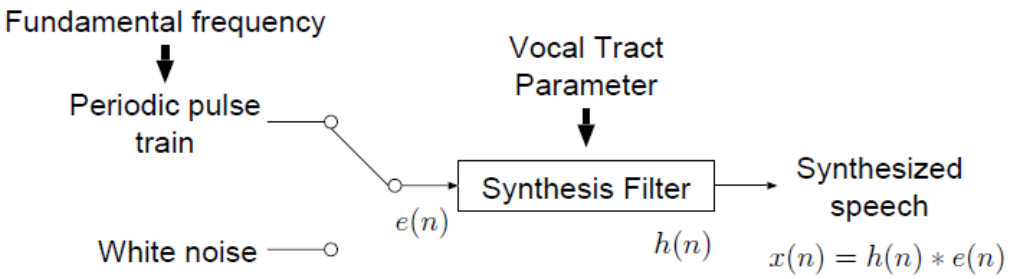
Joonis 1.2: Üksuste valiku skeem [9].

1.3. Mudelipõhised süsteemid

Mudelipõhistes süsteemides ei hoita kõnet, seda kasutatakse vaid mudeli loomise juures. Mudelid koostatakse üksikute kõneüksuste kohta, näiteks kontekstipõhine foneem, ja on indekseeritud lingvistilise spetsifikatsiooni põhjal ning sünteesil moodustatakse järjend sobivatest mudelitest, mille põhjal kõne genereeritakse. Jällegi, antud protseduur ei ole triviaalne, kuna osa mudeleist on puudu, sest kasutada on lõplik hulk treeningmaterjali. Seega peab suutma „lennult“ luua mudeli suvalise soovitud lingvistilise spetsifikatsiooni jaoks. See saavutatakse sidudes parameetrid piisavalt sarnaste mudelitega – analoogne protsess toimub näitepõhiste süsteemide puhul. Seda teemat käsitletakse lähemalt peatükis 2.1.1.

1.4. Signaalitöötlus

Kõnesünteesi signaalitöötlus baseerub üldjuhul allikas-filter mudelil. Vaatleme mudelipõhist süsteemi. Selle konstrueerimiseks on esmalt vaja kõneandmebaasist treeningu käigus ekstraheerida parameetrid, mis kirjeldavad kõnetrakti. Akustilistele tunnustele hinnangu andmiseks kasutatakse mel-sageduse (spekter teisendatakse tajule omasesse sagedusskaalasse [10]) kepstri kordajaid (MFCC – Mel Frequency Cepstral Coefficients).



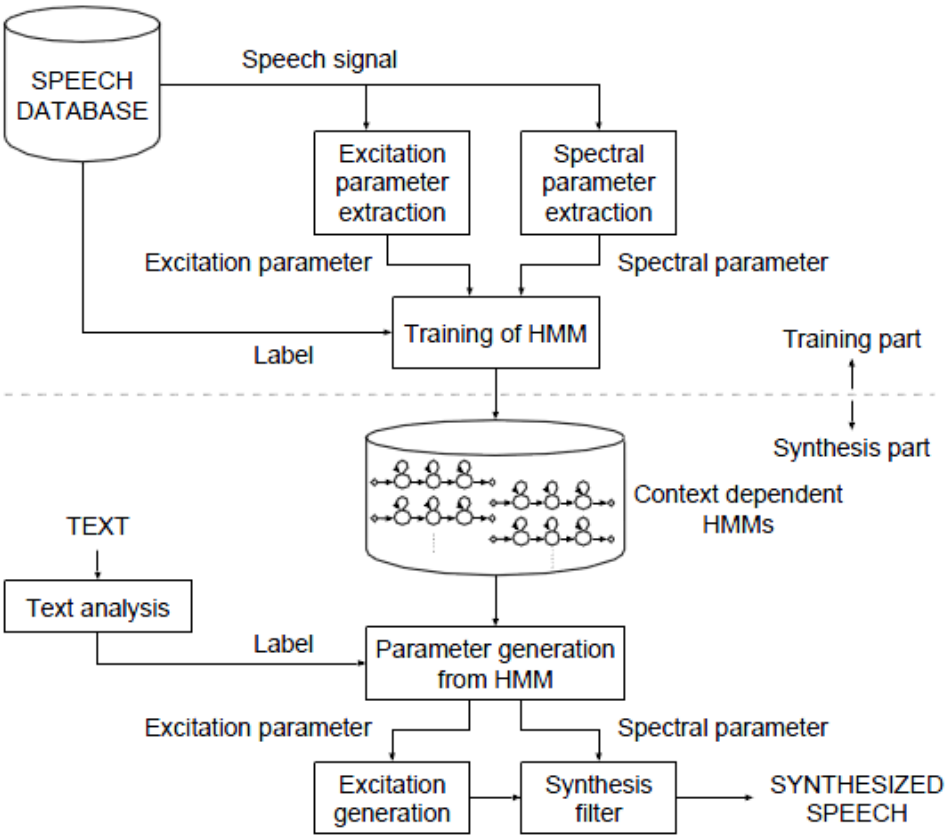
Joonis 1.3: Allikas-filter mudel [11].

Diskreeditud kõnesignaali matemaatilisel käsitlusel kasutatakse allikas-filter mudelit (joonis 1.3). Ülekandefunktsioon *H*(*z*) modelleerib sünteesifiltritena kõnetrakti, hääleallikas on realiseeritud perioodilise helisignaali (helilised häälikud) ja müraallika (helitud häälikud) lülitustena, mis omakorda modelleerib kõne põhitooni. Kõnesignaali *x*(*n*) tootmiseks peavad mudeli parameetrid ajas muutuma. Allikasignaal *e*(*n*) filtreeritakse läbi ajas muutuva lineaarse süsteemi *H*(*z*) ning saadakse kõnesignaal *x*(*n*).

2. Statistiline parameetriline kõnesüntees

Kui räägitakse kõnesünteesi mudelipõhisest lähenemisest, peetakse selle all tavaliselt silmas statistilisi parameetrilisi mudeleid [7]. Mudel on parameetriline, sest kasutab kõnet kirjeldamisel parameetreid, mitte salvestatud näiteid. See on statistiline, sest kirjeldab neid parameetreid statistiliselt: treeningandmetest leitakse parameetrite väärtuste jaotused (näiteks tõenäosuste tiheduse keskmised ja varieeruvused). Süsteemis modelleeritakse Markovi peitmudeleid (HMM – Hidden Markov Model) kasutades üheaegselt helisignaali sagedusspekter (kõnetrakti iseloomustavad tunnused), kõne põhitoon ja foneemide kestused [12].

Markovi peitmudeleid hakati kõnesünteesis kasutama peale seda, kui neid oli edukalt rakendatud kõnetuvastuses. HMM-põhine kõnemudel ei suuda kõnet ülimalt täpselt jäljendada, kuid saadaolevad efektiivsed ja võimekad õppimisalgoritmid (Expectation-Maximization ehk EM-algoritm [13]), automaatsed keerukuskontrollimehhanismid (parameetrite sidumine) ja arvutuslikult efektiivsed otsingualgoritmid (Viterbi algoritm) teevad temast võimsa mudeli.

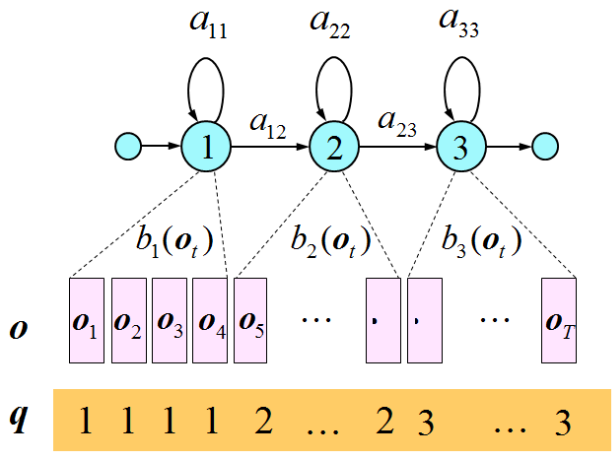


Joonis 2.1: Markovi peitmudelitel põhinev süsteem HTS [9].

2.1. Süsteemi HTS arhitektuur

Joonisel 2.1 on kirjeldatud Markovi peitmudelitel põhineva kõnesünteesisüsteemi HTS[4] toimimine. Süsteem koosneb kahest suuremast osast: treening ja süntees. Treeningfaasis ekstraheeritakse kõneandmebaasist spektriparameetrid, kasutades mel-kepstri analüüsi-tehnikat [14], ja helikõrguse (põhitooni) parameetrid. Saadud infot kasutades treenitakse kontekstipõhised Markovi peitmudelid. Otsustuspuupõhist kontekstiklasterdamise tehnikat [15] kasutades mudelid klasterdatakse ja seotud kontekstipõhistele mudelitele antakse uus hinnang, millega samaaegselt arvutatakse kestusmudelid [16, 17]. Seejärel klasterdatakse kontekstipõhised kestusmudelid, kasutades otsustuspuupõhist klasterdamistehnikat. Sünteesifaasis genereerib lingvistilise töötluse blokk kontekstipõhise märgendi-järjendi, mille alusel konkateneeritakse kontekstipõhised HMM-id ja saadakse lauseahel. Parameetri genereerimise algoritmi kasutades luuakse lausemudelile spektri ja helikõrguse parameetrid, millel rakendatakse seejärel hääleallikat ja sünteesifiltrit genereerimaks sünteeskõne.

Kõnesegmendile vastav Markovi peitmudel on kirjeldatud joonisel 2.2, kus a_{ij} on oleku siirde ja $b_q(o_t)$ väljundtõenäosus ning o on vaatlusjärjend ja q olekujärjend.



Joonis 2.2: Kõnesegmendile vastav Markovi peitmudel [18].

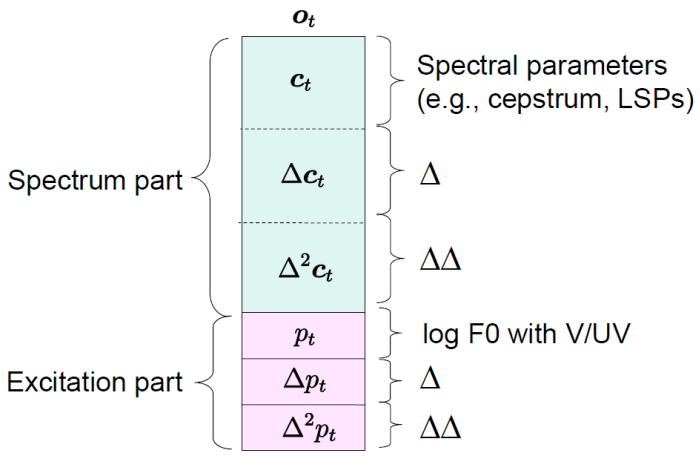
2.1.1. Treeningprotsess

Mudeli parameetritele hinnangu andmisel kasutatakse maksimaalse tõenäosuse kriteeriumi

$$\hat{\lambda} = \arg\max_{\lambda}\{p(O|W,\lambda)\},$$

[4] HTS – akronüüm nimetusest „HMM-based Text-to-Speech“. Süsteemi arendamist alustati Jaapanis Nagoya Tehnikainstituudis 1990-ndate keskel, avatud koodiga tarkvarakomplekt avalikustati 2002. aastal [19].

kus λ on mudeli parameetrite komplekt, O treeningandmete hulk ja W on O-le vastav sõnade järjendite hulk [16]. Hinnangud antakse EM-algoritmi realisatsiooniga [13].



Joonis 2.3: HMM-i väljundvektor [18].

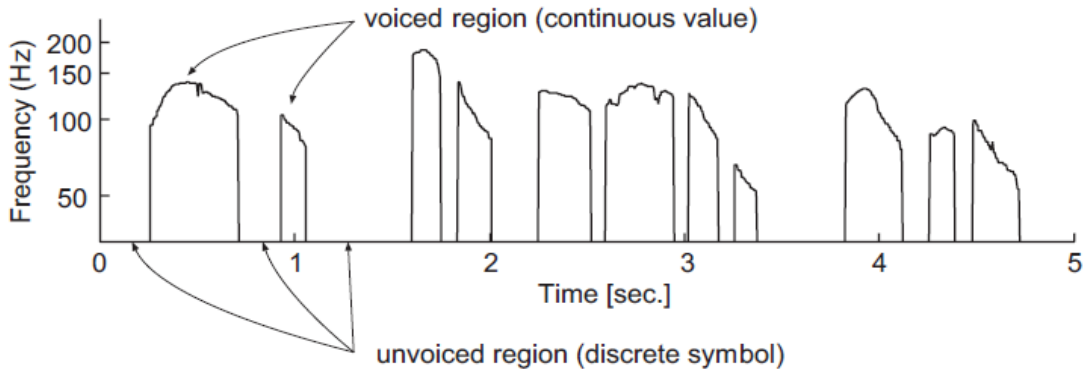
HMM-i väljundvektor o_t (joonis 2.3) koosneb spektri ja helikõrguse osast. Spektriosa sisaldab endas antud juhul mel-kepstri koefitsientide vektorit c_t (sh nullindat järku koefitsient) ja nende delta- ja delta-delta-koefitsiente [11]. Helikõrguse osas kirjeldatakse logaritmitud põhitooniväärtust (log *F0*) ning tema delta- ja delta-delta-koefitsiente, põhitooni mõttes helitut signaali märgitakse diskreetse sümboliga (joonis 2.4). Loomuliku kõlaga kõne loomisel on oluline lisaks staatilistele parameetritele (nt mel-kepstri koefitsiendid ja põhitooni väärtus) modelleerida nende muutumise kiirust. Delta-koefitsient ehk esimest järku tuletis ja delta-delta-koefitsient (teist järku tuletis) arvutatakse antud juhul valemite

$$\Delta c_t = 0{,}5(c_{t+1} - c_{t-1})$$

ja

$$\Delta^2 c_t = c_{t-1} - 2c_t + c_{t+1}$$

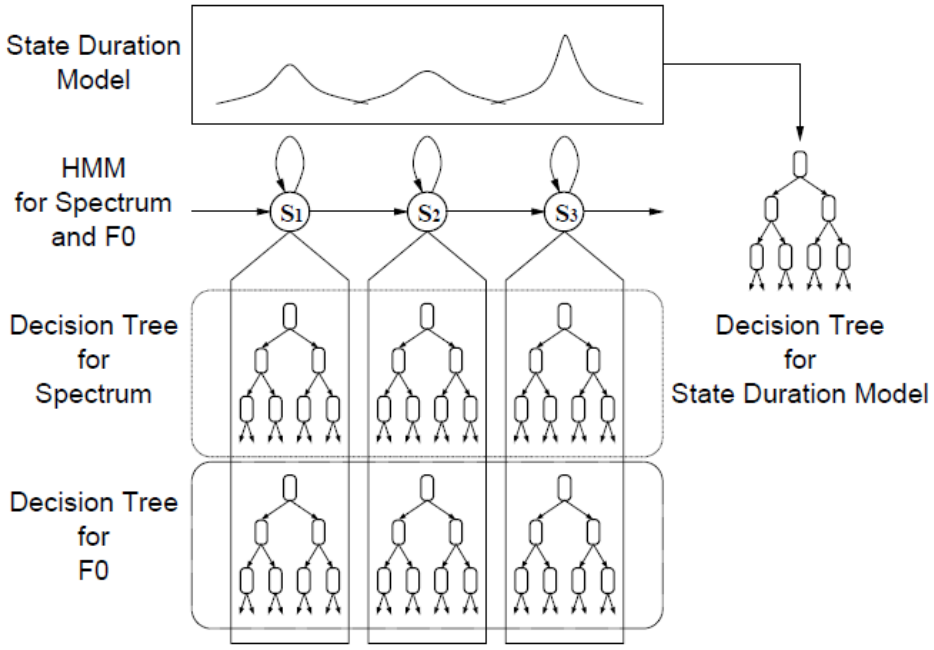
järgi.



Joonis 2.4: Näide põhitoonijärjendist, kus esinevad helilised kui helitud regioonid [16].

Kestusmudeli jaoks jääb Markovi peitmudel liialt lihtsustatuks. Kõrge kvaliteediga sünteeskõne saamiseks modelleeritakse kestus normaaljaotusi kasutades [7, 17]. HMM-ile teise mudeli lisamine teeb temast „Markovi poolpeitmudeli“ (HSMM – Hidden Semi Markov Model): olekutevaheliste üleminekute tasemel on mudel HMM, kuid konkreetse oleku sees enam mitte.

Kõnespektrit, põhitooni mustrit ja kestust mõjutavad mitmed kontekstitegurid (foneemi klassifikatsioon, rõhu ja asukohaga seotud tegurid jm). Mida rohkem erinevaid kontekstitegureid, seda suurem arv kontekstipõhiseid Markovi peitmudeleid, ning kombinatsioonide arv kasvab eksponentsiaalselt, mistõttu ei saa mudeli parameetritele piiratud hulga treeningandmete põhjal täpset hinnangut anda. Lisaks sellele on võimatu koostada kõneandmebaasi, kus kõik kontekstid oleksid kaetud. Probleemi lahendamiseks kasutatakse otsustuspuude põhist kontekstiklasterdamise tehnikat [15]. Klasterdamise tulemusena on erinevate mudelite arv palju väiksem kui erinevate kontekstide arv. Mida suurem on treeningandmete maht, seda suuremat hulka mudeleid saame kasutada. Selle läbi paraneb sünteeskõne kvaliteet, sest süsteemil on võimalik määrata täpsemaid eristusi. Kuna spektrit, põhitooni ja kestust mõjutavad erinevad kontekstitegurid, klasterdatakse need eraldiseisvalt (joonis 2.5). Tulemusena suudab süsteem modelleerida spektrit, helikõrgust ja kestust ühtses raamistikus.



Joonis 2.5: Otsustuspuud kontekstipõhise klasterdamise jaoks [9].

2.1.2. Sünteesiprotsess

Sünteeskõne loomise protsessil analüüsitakse esmalt sisendteksti, mille põhjal luuakse täieliku kontekstiga märgendijada. Selle alusel konstrueeritakse kontekstipõhiseid HMM-e konkateneerides lausemudel. Järgmises faasis luuakse kõneparameetrite genereerimise

algoritmi [20] kasutades lausemudelist kõnespektri ja helikõrguse parameetrid maksimeerides nende väljundtõenäosust valemi

$$\hat{o} = \arg\max_{o}\{p(o|w, \hat{\lambda})\}$$

abil, kus o on kõneparameetrite komplekt, w sünteesitav sõnajärjend ja $\hat{\lambda}$ hinnanguga mudelite hulk [16]. Lõpuks sünteesitakse antud parameetreid ja hääleallikat ning sünteesifiltrit[5] kasutades kõne.

2.2. Statistilise parameetrilise kõnesünteesi eelised

Enamik statistilise parameetrilise kõnesünteesi eelistest näitepõhiste süsteemide ees on seotud tema paindlikkusega, mis tuleneb statistilise modelleerimise protsessist. Kuigi süsteemi treenimine on aeglane (sõltuvalt treeningkorpuse mahust kuni kümneid tunde), ei ole see kriitilise tähtsusega, kuna toimub ainult korra. Kõnemudel on mahult väike ning ei nõua sünteesi realiseerimiseks palju arvutusressursse, mis võimaldab süsteemi kasutada ka väiksema jõudlusega seadmetes.

Kuigi suurem treeningandmete hulk tagab parema sünteesikvaliteedi, on võimalik rahuldava kvaliteediga sünteeskõne saada üksuste valikul põhineva sünteesiga võrreldes palju väiksema mahuga kõnekorpuse pealt, sest statistiline kõnemudel on stabiilne ja suudab segmentaalsetest puudujääkidest hoolimata adekvaatse kõne sünteesida. Arusaadava kõne tootmiseks piisab isegi kümneminutilisest ühe inimese sisseloetud foneetiliselt tasakaalustatud korpusest [21].

Statistilise parameetrilise sünteesi eeliste hulka võib lugeda tema paindlikkuse muuta häälekarakteristikuid, rääkimisstiile ja emotsioone. Kõnelejakohandamiseks vajab süsteem suurusjärku 10 lauset kõnet [22], samuti on võimalik mitme hääledoonori andmetest luua üldine kõnemudel.

Kuna süsteem vajab toimimiseks vaid keelespetsiifilisi kontekstitegureid, on seda võimalik suhteliselt lihtsa vaevaga kohandada soovitud keelele.

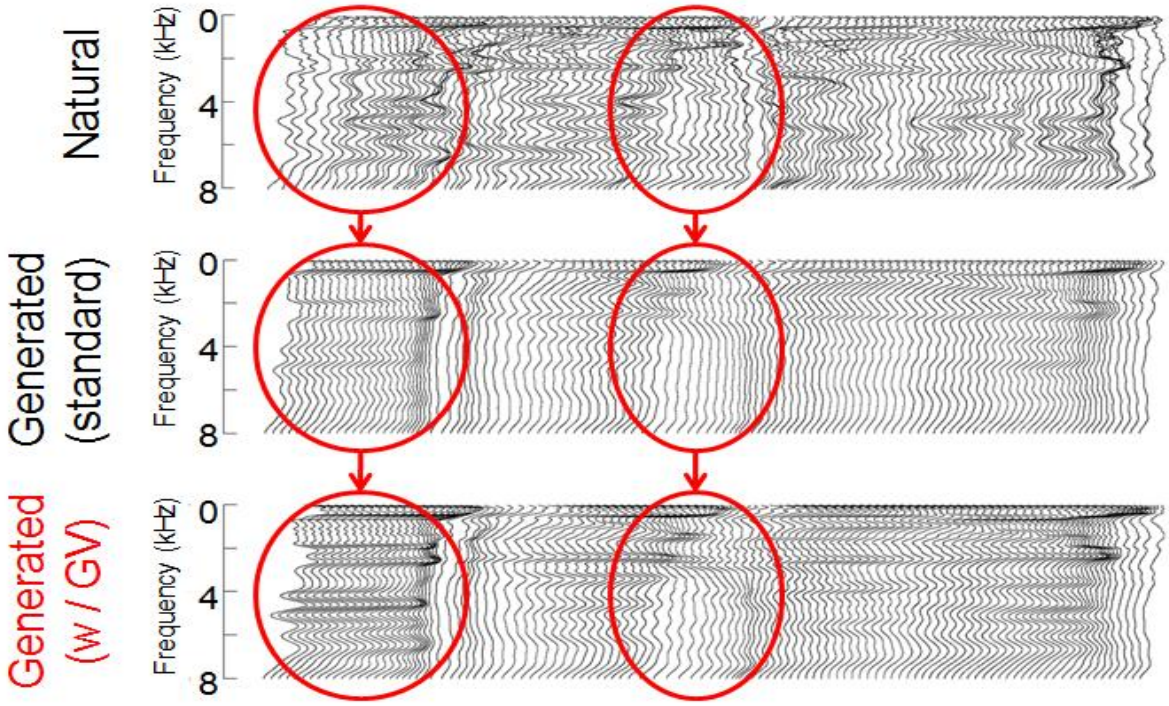
2.3. Statistilise parameetrilise kõnesünteesi puudused

Üksuste valikul põhineva sünteesiga võrreldes peetakse HMM-põhise kõnesünteesi põhiliseks puuduseks genereeritud kõne kvaliteeti. Peamised kitsaskohad on vokooder[6], akustilise modelleerimise täpsus ja liigne silumine [16].

[5] Mel logaritmitud spektri lähendi (MLSA) filter [23].
[6] Vokooder – *vocoder* ehk *voice encoder*, analüüsi-sünteesisüsteem, mis reprodutseerib inimkõnet [24].

Süsteemi HTS sünteeshääl kõlab sumisevana, kuna mel-kepstri vokooder kasutab lihtsat perioodilise helisignaali ja müraallika lülitust. Probleemi lahendamiseks on loodud kõrgekvaliteedilisi vokoodereid, mis võtavad näiteks lisaks arvesse põhitooni aperioodilisust, millega parandatakse segmentide, kus piiri helilisuse ja helituse vahele on mõnel juhul raske tõmmata, kirjeldamist [25]. Eesti keeles on sellisteks foneemideks näiteks „b“, „d“ ja „g“, mis asuvad heliliste häälikute vahel.

Markovi peitmudelid on väga kasulikud erinevate olekute siirete kirjeldamisel, kuid ühe oleku siseselt on info sarnane. Oleku väljundtõenäosus sõltub süsteemis ainult konkreetsest olekust ja oleku kestuse tõenäosus langeb ajas eksponentsiaalselt. Päriskõnes see kahjuks nii ei ole. Seepärast kirjeldatakse näiteks kõneparameetrite trajektoore dünaamiliselt ja kestuse iseloomustamiseks on kasutusele võetud lisamudelid, nagu peatükis 2.1.1 välja toodud.



Joonis 2.6: Globaalse variatiivsuse lisamise mõju sünteeskõne spektrile [18].

Liigne silumine muudab sünteeskõne summutatuks. Statistiline keskmistamine modelleerimisprotsessis muudab süsteemi töökindlamaks andmete hõreduse suhtes, kuid sellega kaasneb liigne silumine, mis kaotab kõnespektrist (formantide) loomuliku variatiivsuse. Probleemi lahendamiseks on kasutusele võetud näiteks globaalse variatiivsuse kirjeldamise tehnika [16], millega suurendatakse sünteeskõne kepstrikoefitsientide liikumisvahemikku. Joonisel 2.6 on näha globaalse variatiivsuse lisamise mõju sünteeskõne spektrile. Välja on toodud loomuliku kõne spekter ning sünteeskõne spekter ilma ja koos globaalse variatiivsusega. Sagedusskaalal nähtavad suuremad võnkumised standardse sünteesisüsteemi genereeritud kõnega võrreldes peegeldavad inimkõne loomulikku spektrivariatiivsust.

3. Kõnekorpus

Süsteemi HTS toimimise eelduseks on foneemi tasandil kontekstuaalselt märgendatud kõnekorpuse olemasolu. Kõnekorpus on antud juhul vajalik kõnemudeli treenimisel. Kuigi statistilise parameetrilise kõnesünteesi toimimiseks piisab võrdlemisi väikesest treeningkorpusest, tõstab kõrgekvaliteediline suure mahuga andmehulk märgatavalt sünteeskõne kvaliteeti.

Käesolevas töös on kasutatud Eesti Keele Instituudis salvestatud kõnekorpust, mille kogumaht on ca 17 tundi helistuudios salvestatud kõnet kokku viielt kõnejuhilt (kolm raadiodiktorit, kaks kõnesünteesi arendamisega seotud inimest), igaühelt 700-3500 loetud lauset. Korpuse loomine on koos lingvistilise töötluse arendamise, erinevate sünteesimeetodite kasutamise ja kõnematerjalide hindamisega iteratiivne protsess.

3.1. Korpuse loomine

Esimene kõnekorpus eestikeelse kõnesünteesi jaoks salvestati 1998. aastal difoonsünteesi tarbeks ning see sisaldab ca 1700 difooni [26]. 2006. aastal käivitatud projekti „Eestikeelne korpuspõhine kõnesüntees“ [1, 27] raames loodi kõnekorpus üksuste valikul põhineva kõnesünteesi jaoks. Esmane eesmärk oli luua korpus, milles oleks ca 60 minutit kõnematerjali, kuid foneetilisest ja fonoloogilisest aspektist piisavalt esinduslik, sisaldades kõiki kõnes esineda võivaid difoone, palju numbreid ja aastaarve, tihedamini esinevaid sõnu, väljendeid ja morfoloogilisi tunnuseid [28]. Süntaktilisest aspektist püüti sõnad asetada konstrueeritud lauses võimalikult loomulikku konteksti, leides Leipzigi Ülikooli korpusteportaali Wortschatz [29] abil nende tihedamini esinevad kollokatsioonid. Esmane korpus sisaldas endas 400 lauset (ca 50 minutit kõnematerjali).

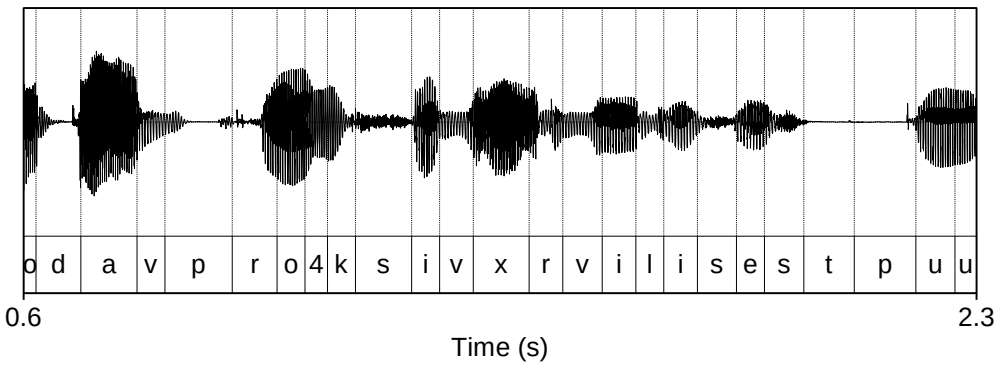
Järgmistes faasides on püütud kõnekorpuse foneetilist rikkust suurendada (lisati näiteks diftonge erinevates positsioonides), konstrueeritud lausete kõrval on kasutusele võetud ka ilukirjanduslikke tekste ja ajalehtedest uudiseid. Jälgitud on lausete pikkust (mitte üle kolme osalause) ja neutraalsust (ilukirjanduslikke lauseid püüti lugeda emotsioonitult), hoidutud kasutamast võõrnimesid (võimaliku ebasobiva prosoodia tõttu).

Kuna üksuste valikul põhineva kõnesünteesi puhul on kõnematerjali kvaliteet kriitilise tähtsusega, sobib sellel eesmärgil loodud korpus statistilise parameetrilise sünteesi jaoks väga hästi.

3.2. Lingvistiline töötlus

Eestikeelsel korpuspõhisel kõnesünteesil kasutatakse selleks spetsiaalselt loodud lingvistilise töötluse vahendeid. Samu mooduleid on kasutatud ka käesoleva töö raames (süsteemi HTS tekstianalüüsil), kuna eesmärgiks on luua häälemudelid, mida saab kaasata eestikeelse kõnesünteesi rakendustesse.

Korpuse esialgsel salvestamisel, märgendamisel ja segmenteerimisel joonduti eesti keele kanoonilise häälduse järgi [28], mille aluseks on Eesti Õigekeelsussõnaraamat [30]. Lingvistilise töötluse võimekuse arenedes võeti paralleelselt kasutusele automaatse segmenteerimise rakendused ning edaspidi on märgendatud kõnekorpus vastavalt tekstianalüüsi väljundile. Joonisel 3.1 on toodud näide kõnekorpuse foneetilisest transkriptsioonist ja segmenteerimisest.



Joonis 3.1: Foneetilise transkriptsiooniga varustatud lõik lausest „Üks odav pronksivärvilisest puust lühter, ...“. Näide kõnejuht Liisilt. Automaatselt segmenteeritud.

Kuna lingvistilises töötluses esineb vigu (näiteks kolmanda välte äratundmise täpsus on ca 90%), ei ole teksti analüüsil garanteeritud korrektne transkriptsioon, mis vähendab sünteesi kvaliteeti [31]. Samas ei ole mõtet konstrueerida ideaalset kõneandmebaasi, kui tekstianalüüsiga ei suudeta korpuses esinevaid segmente „välja kutsuda“. Võib treenida Markovi peitmudelitega ideaalsel või ideaalilähedasel korpusel kõnemudeli, kuid kõne sünteesimisel ei ole paljudest mudelitest ja klastritest kasu, kuna lingvistiline töötlusblokk genereerib ideaalse asemel puuduliku hinnangu. Seega on mõistlik kasutada häälemudeli treenimisel korpust, mis vastab tekstianalüüsi võimekusele.

3.3. Kõnematerjalide salvestus ja hindamine

Hääledoonori valikul on peamisteks kriteeriumiteks võimekus teksti lugeda korrektse häälduse ja ajas suhteliselt sarnaste prosoodiliste parameetritega [28]. Seetõttu on korpuse kõnejuhtideks valitud raadiodiktorid ja lingvistilist tausta tundvad (antud juhul ka

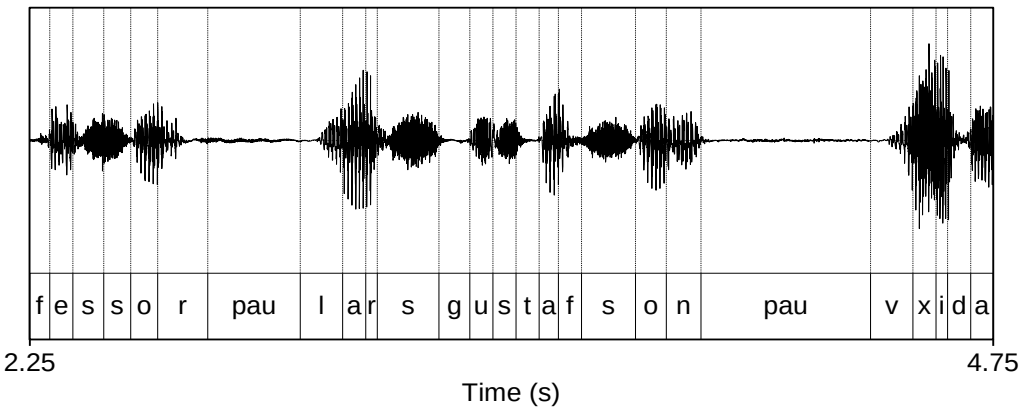
kõnesünteesiga seotud) inimesed. Tabelis 3.1 on välja toodud kõnekorpuse orienteeruv maht kõnejuhi kaupa.

Tabel 3.1: Kõnekorpuse orienteeruv maht kõnejuhi kaupa.

Kõnejuht	Raadiodiktor	Lauseid	Maht tundides
Riina	+	700	1
Einar	+	1000	1,5
Tõnu	+	3000	6
Liisi		3400	6,5
Tõnis		1500	1,5

Kvaliteetse kõnekorpuse saamiseks on salvestised tehtud helistuudios. Helimaterjal on seejärel käsitsi lausestatud ning vigase häälduse laused korpusest eemaldatud. Järgnevalt on korpus automaatselt segmenteeritud tekstianalüüsi väljundiks saadud transkriptsiooni põhjal. Segmenteerimise tulemus on kontrollitud kuulamise teel üle ning lisatud foneemitasandile vajalikesse kohta pausid, mida tekstianalüüs või segmenteerija ei ole suutnud ennustada. Sõnade transkriptsiooni ei ole muudetud, olulise märgendusvea korral on lause korpusest eemaldatud.

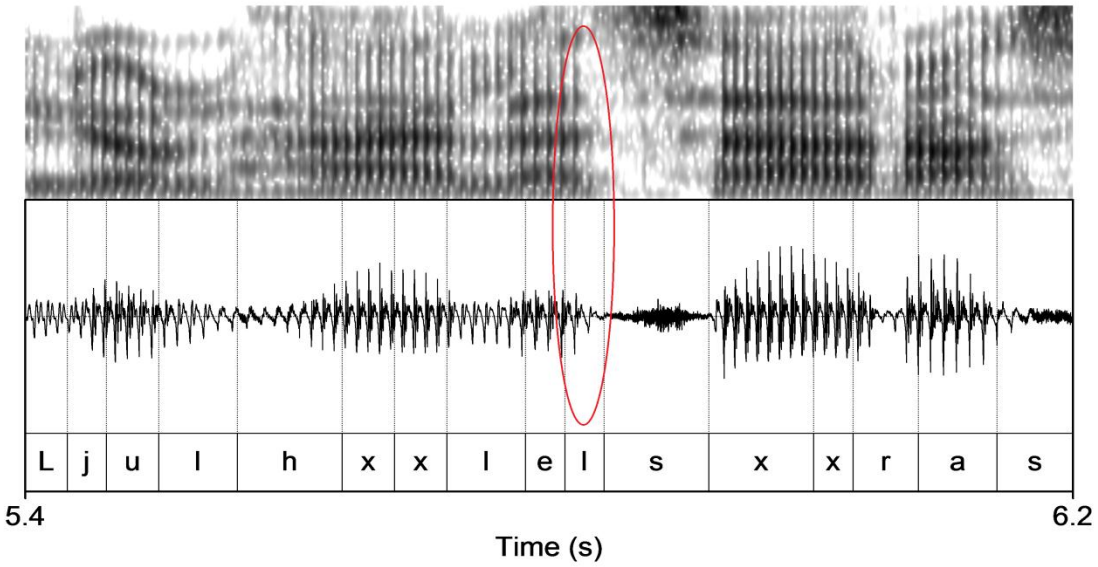
Pauside lisamine on vajalik, et kõnesegmendi koosseisu ei sattuks lugemisel tehtud hingamis-, rõhutus-, või mõttepause. On ilmnenud, et näiteks pärisnimed öeldakse tihti välja rõhutatult ja tehakse nende ette paus. Seetõttu on edaspidi korpuse koostamisel nimest välja jäetud perekonnanimi või on nimi asendatud neutraalsega (näiteks Jüri või Mari). Joonisel 3.2 on näha, kuidas kõnejuht on lause „Islandi ülikooli majandusprofessor Lars Gustafson väidab, ...“ lugemisel pärisnime rõhutades teinud selle ette ja järele pausi.



Joonis 3.2: Lõik lausest „Islandi ülikooli majandusprofessor Lars Gustafson väidab, ...“. Näide kõnejuht Einarilt. Automaatselt segmenteeritud.

Kuigi kõnejuhid on näiliselt hoolikalt valitud, on esinenud ka selles aspektis tagasilööke. Näiteks hääledoonor Tõnu puhul ilmnes kõnelainet uurides, et kõrvale tajutava väga selge ja kiire kõne kiirus on saavutatud mitmete foneemide ütlemata jätmisega. Joonisel 3.3 on

välja toodud lõik Tõnu loetud fraasist „nad loopisid valjul häälel sääraseid sõnu“, kus on sõnas „häälel“ näha häälikukadu viimase foneemi „l“ puhul: spektrogrammi (joonise ülemine kolmandik) uurides on tuvastatud üleminek foneemilt „e“ foneemile „s“ ning kõnelaineski ei sarnane „l“ esitus teistele samal joonisel leiduvatele foneemidele „l“. Seetõttu on üksuste valikul põhineva sünteesi puhul saanud Riina sünteeshääl Tõnu omast paremaid hinnanguid, kuigi Riina puhul on kasutada kordades vähem materjali.



Joonis 3.3: Häälikukadu fraasis „nad loopisid valjul häälel sääraseid sõnu“. Näide kõnejuht Tõnult. Automaatselt segmenteeritud.

Andmemahtude suhtelisest piiratusest tulenevalt on kasutatav alamkorpus mitu korda hoolikalt üle hinnatud ja nii on mitmel juhul tehtud kompromisse, kaasates vähem kvaliteetseid lauseid. Erinevate katsetuste käigus on siiski ühe kõnejuhi (Tõnu) loetud materjalist välja jäetud kuni veerand lauseist.

4. Kõnemudeli loomine

Käesoleva töö põhieesmärgiks on luua süsteemiga HTS Markovi peitmudelitel põhinev häälemudelid mees- ja naishäälele, mida on võimalik kasutada eestikeelse kõnesünteesi rakendustes. Selleks on vaja kohandada HTS eesti keelele ning ühildada väljundmudel tekstianalüüsiga. Süsteemi kohandamine eesti keelele tähendab sisuliselt foneetilise ning fonoloogilise spetsifikatsiooni koostamist ja treeningmaterjalide ettevalmistamist. Eestikeelse kõnesünteesi rakendustes kasutamiseks peab antud spetsifikatsioon ühtima lingvistilise töötluse moodulitega. Mudelitele antakse esmane hinnang nende põhjal genereeritud näitelausete alusel.

4.1. Süsteemi HTS kohandamine eesti keelele

Süsteemi HTS kodulehelt allalaaditav treeningdemo [32] on koostatud mudeli loomiseks inglise keelele. Demo koosneb andmefailidest ja skriptidest. Andmefailide hulka kuuluvad helifailid (raw-formaadis), neile vastavad kontekstuaalset infot (foneetiline märgendus, segmenteerimine jm) sisaldavad lausungifailid, foneetilisi ja fonoloogilisi kontekstikategooriaid sisaldav infofail ning soovi korral HTS-iga automaatselt näitelausete genereerimiseks süsteemi HTK (Hidden Markov Model Toolkit) [33] märgendusformaadis kontekstiinfot sisaldavad lausungifailid. Helifailidele vastavad lausungifailid on antud juhul süsteemi Festival utt-formaadis, mis tagab HTS-i loodud mudeli ühilduvuse Festivali eeskomponendiga. Samas ei kasutata häälemudeli treenimisel utt-failis sisalduvat foneemipiiride infot, sest kõik Markovi peitmudelite parameetrid määratakse treeningprotsessi käigus automaatselt [9].

4.1.1. Foneetiliste ja fonoloogiliste kontekstide moodustamine

Kontekstikategooriad võib jagada kahte ossa: foneetilised ja fonoloogilised. Esimene hõlmab endas foneemide klassifikatsiooni, teine fonoloogilisi kontekstifaktoreid. Käesolevas töös on fonoloogiliste kontekstifaktorite kirjeldamisel võetud aluseks süsteemi HTS demoversioonis pakutud tegurid inglise keelele (tabel 4.1) ning muudetud need sobivamaks eesti keelele. Foneetikakirjeldus on kohandatud vastavalt spetsifikatsioonile, mida kasutatakse süsteemis Festival eestikeelse kõnesünteesi tekstianalüüsil.

Tabel 4.1: Inglise keelemudelil põhinevad fonoloogilised kontekstifaktorid tasemete kaupa [9].

Foneem	
	{eelnev, praegune, järgnev} foneem
	praeguse foneemi asukoht silbis
Silp	
	foneemide arv {eelnevas, praeguses, järgnevas} silbis
	{eelneva, praeguse, järgneva} silbi {aktsent, rõhk}
	silbi asukoht sõnas
	{eelnevate, järgnevate} {rõhuliste, aktsendiga} silpide arv fraasis
	silpide arv {eelmise, järgmise} {rõhulise, aktsendiga} silbini
	silbis asuv vokaal
Sõna	
	{eelneva, praeguse, järgneva} sõna leksikaalse klassi hinnang
	silpide arv {eelnevas, praeguses, järgnevas} sõnas
	sõna asukoht fraasis
	{eelnevate, järgnevate} täistähenduslike sõnade arv fraasis
	sõnade arv {eelneva, järgmise} täistähendusliku sõnani
Fraas	
	silpide arv {eelnevas, praeguses, järgnevas} fraasis
	fraasi lõputoon
Lausung	
	silpide arv lausungis

Eesti keeles on traditsiooniliselt 26 kvalitatiivselt erinevat segmentaalfoneemi [8]: 9 vokaali *a, e, i, o, u, õ, ä, ö, ü* ja 17 konsonanti *f, h, j, k, l, l′, m, n, n′, p, r, s, s′, š, t, t′, v* (*l, n′, t′* ja *s′* on palataliseeritud). Kõnesünteesis on lisaks neile kasutusel ninahäälik *ŋ*[7] ja poolvokaal *w*[8], samuti erinevad vaikust tähistavad märgendid.

Foneetikakirjeldusi on eestikeelse korpuspõhise kõnesünteesi arendamise käigus olnud kasutusel mitmeid [28, 34]. Peamised variatsioonid on esinenud foneemide pikkuste ja väldete märkimisel ning klusiilide esitusel. Nii näiteks on silbipiiril heliliste häälikute vaheline klusiil märgendatud ühes versioonis ühe-, teises kahekordsena (sõna „kata“ transkriptsioon *kata* või *katta*). Foneemi *a* erinevaid realisatsioone on kirjeldatud näiteks kahel kuni neljal viisil (tabel 4.2). Pikk *a* on märgendatud ja segmenteeritud esimese esitusviisi puhul kahe foneemina *a* ja *a*, teise esitusviisi puhul ühe foneemina *ax*.

[7] „g“-le ja „k“-le eelnev „n“ hääldatakse *ŋ*-na, välja arvatud liidete -gi ja -ki puhul.

[8] Silbipiiril pika või ülipika „u" või „u“-ga lõppeva diftongi järel ja järgneva vokaali ees hääldatav poolvokaal. Näiteks sõna „kaua“ foneetiline transkriptsioon on *kauwa*.

Tabel 4.2: Foneemi *a* esitusviise kõnesünteesi erinevates arendusettapides.

Lühike "a" (sõnas "sadu")	Pikk "aa" (sõnas "saamid")	Ülipikk "a", ühekordne (sõnas "pea")	Ülipikk "a", kahekordne (sõnas "maa")	Esitus 1	Esitus 2
+				a	a
	+			a + a	ax
		+		a:	ay
			+	a + a:	az

Niisamuti on foneetikakirjeldustega eksperimenteeritud Markovi peitmudelitel põhineva häälemudeli loomisel. Foneemide kirjeldamise võimalused kombineerituna foneemide arvuga annavad potentsiaalselt suure hulga erinevaid märgendeid transkriptsiooni jaoks. Tabelis 4.2 toodud teise esitusviisi puhul on see arv suurusjärku 100. Teise äärmusena on katsetatud mudeli loomist foneemide esitusviisiga, kus on kasutatud vaid „lühikesi“ foneeme ehk pikkuse mõõde on kõrvale jäetud ning saadud minimaalne hulk foneeme, mis eesti keele kvalitatiivselt katavad.

Nagu peatükis 2.1.1 välja toodud, ei suuda süsteem väga suure kontekstide arvu puhul treeningprotsessi käigus mudeli parameetritele suure tõenäosusega adekvaatset hinnangut anda. Seega ei ole mõistlik kasutada väga suurt hulka erinevaid foneeme. Samas genereeritakse väikese foneemide hulga peal liialt lihtsustatud mudel. Optimaalne erinevate foneemide hulk võib olla orienteeruvalt 50. Süsteemi HTS kasutaval kõnesünteesil on inglise keele puhul kasutusel 51 [32], rootsi keelel 53 [35] ja katalaani keelel 38 foneemi [36].

Mudeli treeninguks vajalikus infofailis on lisaks fonoloogilistele kontekstidele kirjeldatud foneetilised klassid ehk foneemid on jaotatud neile iseloomulike tunnuste põhjal kategooriatesse. Nii kuuluvad lisas 2 toodud häälemudeli „Liisi“ foneetikakirjelduses vokaalide klassi lühikesed *a, e, i, o, u, q* (ehk *õ*), *ae* (ehk *ä*), *c* (ehk *ö*), *y* (ehk *ü*), ja pikad *a:, e:, i:, o:, u:, q:, ae:, c:, y:*. Poolvokaalide hulka on määratud *j, j:* ja *w* ning näiteks helitute hõõrdhäälikute (*unvoiced fricative*) klassi *f, s, S* (ehk palataliseeritud *s*), *sh* (ehk *š*), *h, f:, s:, S:, sh:* ja *h:*.

4.1.2. Treeningkorpuse valimine

Treeningkorpuse valimisel kehtib reegel: mida rohkem andmeid, seda täpsem mudel on võimalik luua. Käesoleva töö raames on eksperimenteeritud Riina, Liisi ja Tõnu alamkorpustega, millest on moodustatud erineva mahuga treeningkorpusi, kusjuures arvesse on võetud foneetilise tasakaalustatuse aspekti, mille kohaselt on oluline kõikide foneemide esinemine võimalikes kontekstides (fraasi algus ja lõpp, rõhuline ja rõhutu positsioon, heliline ja helitu ümbrus jne) soovitavalt mitu korda.

Katseid on tehtud ka väga väikesemahuliste treeningkorpustega, leidmaks minimaalset andmemahtu, mis on vajalik arusaadava sünteeskõne genereerimiseks. Kuna see ei ole olnud eesmärk omaette, on treeningandmete, kontekstifaktorite ja tekstianalüüsi komplekt optimeerimata ning nii häid tulemusi, milleni jaapanlastel jõudsid – arusaadav sünteeskõne kümneminutilise treeningkorpuse pealt [21] – ei ole õnnestunud saada.

4.2. Treenitud häälemudelite hindamine

Süsteemiga HTS treenitud häälemudelitele on hinnang antud peamiselt nende põhjal genereeritud näitelausete põhjal. Näitelausete komplekt koosneb 150 lausest, kusjuures on välditud nende lausete sattumist treeningkorpusesse. Eksperimentidel Riina ja Liisi alamkorpustega on kasutatud vanemat ja algelisemat tekstianalüüsikomponenti, Tõnu alamkorpusel treenitud häälemudeli juures on kasutusel uuem parendatud eeskomponent.

Valiku näitelausetest võib leida antud tööga kaasa pandud CD-plaadilt (lisa 3), millel on välja toodud kõnejuht Liisi erinevatel treeningkorpustel ja lingvistilistel spetsifikatsioonidel loodud häälemudeli põhjal genereeritud näitelaused ning võrdluseks häälemudeliga „Tõnu“ sünteesitud laused.

Liisi alamkorpusest moodustati treeningkorpused mahuga 100, 250, 500, 1000 ja 2000 lauset. Häälemudelite treenimisel kasutati esmalt lisas 1 toodud foneetilist tähestikku, seejärel kasutati vaid „lühikesi“ ehk kvalitatiivfoneeme[9].

Oodatult saadi kõige parem tulemus valitutest suurima, 2000-lauselise korpusega. 100-lauselise korpuse põhjal loodud häälemudel aga ei olnud piisavalt hea, et sellega arusaadavat sünteeskõnet tekitada. Helinäited on toodud CD-plaadil kaustas `liisi_100/`. 250-lauselisel korpusel treenitud mudeliga genereeritud sünteeskõne on arusaadav. Näitelaused nimega `liisi_lyh_250.wav` asuvad kaustades `näitelause_1/ .. näitelause_8/`.

500-lauselisel korpusel treenitud mudelile võib sünteeskõne arusaadavuse aspektist anda rahuldava hinnangu, kuid leidub kohti, kus häälemudel on puudulik. Näiteks sünteesitud lause „Lambda on kreeka täht“ (CD-plaadil fail `liisi_500/test0390.wav`) sisaldab kohati arusaamatut müra.

Ainult kvalitatiivfoneemide kasutamine tähestikus kahandab mõningal määral sünteeskõne kvaliteeti, kui tegemist on keskmise suurusega (500 lauset) või suurema korpusega, kuna statistiline keskmistamise protsess vähendab erinevatele parameetritele hinnangu andmisel kõnele iseloomuliku variatiivsuse peegeldamist. Väiksema korpuse puhul on efekt pigem

[9] Kvalitatiivfoneemide esitus süsteemi HTS häälemudeli „Liisi“ puhul on näha lisas 1 tabeli veerus „Lühike“.

vastupidine, sest väiksem kontekstitegurite hulk võimaldab niigi puudulike andmete pealt suurema tõenäosusega adekvaatse mudeli luua.

Sünteesitud kõne kvaliteet on hea, kui ta on arusaadav ning selles ei ebaloomulikke hälbeid häälikute kvaliteedis ega häälduslvigasid, mille seast hakkavad eriti kõrva vead välte määramisel. Näitelausete võrdlemisel on antud hinnang, et kõige olulisem tegur häälemudeli kvaliteedi juures on kvaliteetne treeningkorpus ehk kontekstuaalse info ekstraheerimiseks piisava materjali olemasolu. Andmemahtude suurendamisel häälemudeli kvaliteedi tõus aeglustub märgatavalt. Järgmine määrav komponent on tekstianalüüs. Kui sünteeskõne on arusaadavuse aspektist heal tasemel, mõjutavad sellele antud hinnangut negatiivselt eelkõige vead, mis on tehtud ortograafilise teksti teisendamisel häälduste kstiks. Neist vähem, kuid siiski olulisel määral, avaldavad mõju valitud foneetilised ja fonoloogilised kontekstitegurid.

Kaasasoleval CD-plaadil on võimalik võrrelda kaustades `näitelause_1/ .. näitelause_8/` leiduvaid erinevate häälemudelitega genereeritud näitelauseid ning anda eelmises lõigus toodud printsiipidele omapoolne hinnang. Ära on toodud sünteesitud laused häälemudeliga „Tõnu“ ning kõnejuht Liisi erineva mahuga treeningkorpustel loodud mudelit kasutades genereeritud laused. Faili nimes sisalduv indikaator `lyh` viitab mudeli treenimisel ainult kvalitatiivfoneemide kasutamisele.

4.3. Häälemudeli ühildamine Festivali eeskomponendiga

Viimastel aastatel on süsteemide Festival ja HTS arendajad palju koostööd teinud ning jõutud on tulemuseni, kus HTS-iga treenitud häälemudel on suhteliselt lihtsa vaevaga võimalik ühendada Festivali eeskomponendiga. Ühilduvuse põhiliseks eelduseks on häälemudeli treeningul tekstianalüüsi võimekusele vastava materjali kasutamine ehk lingvistiliste spetsifikatsioonide kokkulangevus. Ühildamise protseduuri täpsem kirjeldus jääb väljapoole antud töö eesmärgipüstitust.

4.4. Järgnevad võimalikud tegevused sünteesikvaliteedi tõstmiseks

Kui välja jätta eeskomponendi arendustööd, võib sünteesikvaliteedi tõstmiseks teha järgnevat: võtta kasutusele kvaliteetsem vokooder (süsteem, mis mudeli põhjal reprodutseerib inimkõnet) ning optimeerida foneetilisi ja fonoloogilisi kontekstitegureid.

Kvaliteetsema vokooderi kasutuselevõtt vähendab suminat sünteeskõnes. Levinud on kõrgekvaliteediline vokooder STRAIGHT [37], mis kasutab lisaks hääle- ja müraallikale sünteesifiltri juures ka helilaine aperioodilisuse tunnuseid, kirjeldamaks helitu ja helilise spektrikomponendi koosesinemist [7, 25].

Foneetiliste kontekstitegurite hulga suurendamine kirjelduse täpsuse eesmärgil suurendab eksponentsiaalselt kontekstuaalsete mudelite arvu, mis on vajalikud kõnemudeli kirjeldamiseks. Seega on oluline foneemide kvantitatiivsete esituste hulga optimeerimine (kas näiteks hääliku „f“ esitusel on vaja eristada foneeme *f* ja *f:*).

Kõnemudeli treenimisel kasutatavad fonoloogilised tegurid ei ole täielikult eesti keeles kirjeldatud tingimustele vastavad. Samuti on kontekstuaalseid tegureid, mille mõju eesti keelele ei ole uuritud, kuid mis on andnud mõne teise keele puhul häid tulemusi. Sellest tulenevalt võib arvata, et antud temaatikasse süvenedes on võimalik eestikeelse statistilise parameetrilise kõnesünteesi kvaliteeti tõsta.

Kokkuvõte

Antud bakalaureusetöös anti ülevaate Markovi peitmudelitel põhineva häälemudeli loomisest eestikeelse kõnesünteesi rakenduste jaoks.

Esmalt tutvustati tekst-kõne sünteesi protsessi, kirjeldati tüüpilise sünteesisüsteemi komponente ning vaadeldi enamlevinud paradigmade lähenemist kõnesünteesile.

Täpsemalt käsitleti statistilist parameetrilist kõnesünteesi ja selgitati antud töö raames kasutatud Markovi peitmudelitel põhineva sünteesisüsteemi HTS toimimismehhanisme, anti ülevaade tema eelistest ja puudustest ning võimalikest probleemilahendustest.

Praktilises osas kasutati Eesti Keele Instituudis koostatud ja salvestatud kõnekorpust. Välja toodi korpuse loomise põhimõtted ning seos kõnesünteesisüsteemi lingvistilise töötluse mooduliga ning sellest tulenevad piirangud. Kirjeldati tekstianalüüsi arendamisega kaasnenud muutusi häälikusüsteemi valikul. Ära märgiti kõnekorpuse salvestamisega seotud aspektid ja materjalide hindamise põhimõtted ning analüüsiti korpuse kvaliteeti mõjutanud leide, millest tulenevalt on muudetud järgnevate korpuste koostamise põhimõtteid.

Töö eesmärgiks olnud häälemudeli loomisel toodi esmalt välja süsteemi HTS kohandamine eesti keelele, mis sisuliselt tähendas foneetilise ja fonoloogilise spetsifikatsiooni koostamist ja treeningmaterjalide ettevalmistamist. Kuna sooviti võtta häälemudel kasutusele eestikeelse kõnesünteesi rakendustes, tuli spetsifikatsioon ühildada saadaval oleva tekstianalüüsi omaga.

Katseid tehti erinevate kõnejuhtide erinevate alamkorpustega ja eksperimenteeriti lingvistilise spetsifikatsiooniga. Välja toodi mees- ja naishäälele treenitud mudelitega genereeritud sünteeskõne näited, mille põhjal anti ka hinnang mudelite headusele.

Ootuspärase tulemusena leiti, et olulisimad tegurid häälemudeli kvaliteedi juures on treeningkorpuse maht ja kvaliteet. Teine määrav komponent on tekstianalüüs ja tema võimekus efektiivselt teisendada ortograafiline tekst häältdustekstiks. Olulisuselt kolmandaks headuse hinnangu mõjutajaks hinnati foneetiliste ja fonoloogiliste kontekstitegurite optimeerimine.

Lõpuks toodi ära võimalikud tegevused, mille tulemusena on võimalik Markovi peitmudelitel põhineva kõnemudeliga genereeritud sünteeskõne kvaliteeti tõsta.

Creation of HMM-based Speech Model for Estonian Text-to-Speech Synthesis

Bachelor thesis

Tõnis Nurk

Summary

The main purpose of this thesis is to create hidden Markov model based speech models for both male and female voice for Estonian text-to-speech synthesis.

To begin with, a brief overview of text-to-speech synthesis process is given, alongside with description of components in a typical speech synthesis system and popular techniques in common use.

Subsequently, the thesis focuses on statistical parametric speech synthesis in particular. The technique called hidden Markov model-based speech synthesis which is utilized in the system HTS (HMM-based Speech Synthesis System) is described. HTS is employed to generate voice models needed for this bachelor work. Discussed are the advantages and drawbacks of the system HTS and described are solutions to some of the problems.

In the practical part of the work the creation of speech corpus in Institute of the Estonian Language is analyzed. Presented are the guidelines for creation of the corpus as well as its connection with text analysis module and related constraints. Described are the changes to phonetic system in use followed from development of text analysis modules. Given are the aspects related to recording the speech corpus and guidelines to evaluate the quality of the signal produced. Analyzed are the unforeseen findings that affect quality of the corpus and from these new guidelines for corpus construction are derived.

Described is the process of adapting Estonian-related training data and linguistic specification to the system HTS. Linguistic specification is compatible with text analysis module in order to enable implementation of the trained voice models to Estonian speech synthesis applications.

Experiments are carried out on data from different speakers, subcorpora and linguistic specifications. Presented are examples of generated speech for both male and female voice models trained with HTS.

Speech model evaluation process has given expected findings. The most important factors that affect voice model quality are the quality and size of training corpus. It is followed by

the ability of text analysis module to generate accurate pronounciation text and optimizing of phonetical and phonological contextual factors.

In the end, proposed are two possible courses of action to improve the quality of HMM-based speech models trained: implementation of STRAIGHT vocoder to reduce buzzyness of synthesized speech and optimizing of phonetical and phonological contextual factors.

Viited

1. M. Mihkla, I Hein, I. Kiissel, T. Nurk, L. Piits. *Eesti keele tekst-kõne süntees*, http://www.eki.ee/keeletehnoloogia/projektid/syntees/tks.html [Viimati külastatud 10.05.2012]

2. E. Meister, L. Meister, R. Metsvahi. *Audiovisuaalse kõnesünteesi prototüüp*, http://www.phon.ioc.ee/dokuwiki/doku.php?id=projektid:avsyntees:avsyntees.et [Viimati külastatud 10.05.2012]

3. T. Masuko, K. Tokuda, T. Kobayashi, S. Imai. *Speech synthesis from HMMs using dynamic features.* Proceedings of ICASSP, 1, lk 389-392, 1996.

4. K. Tokuda, K. Oura, K. Hashimoto, S. Shiota, H. Zen, J. Yamagishi, T. Toda, T. Nose, S. Sako, A. W. Black. *HMM-based Speech Synthesis System (HTS)*, http://hts.sp.nitech.ac.jp/ [Viimati külastatud 10.05.2012]

5. A. W. Black, R. Clark, K. Richmond, J. Yamagishi, V. Strom, S. King. *The Festival Speech Synthesis System*, http://www.cstr.ed.ac.uk/projects/festival/ [Viimati külastatud 10.05.2012]

6. P. Taylor. *Text-to-Speech Synthesis*. Cambridge University Press, 2009.

7. S. King. *An introduction to statistical parametric speech synthesis*, Sādhanā, 36(5), lk 837-852, 2011.

8. T. Erelt, M. Erelt, K. Ross. *Eesti keele käsiraamat*, Eesti Keele Sihtasutus, 2007.

9. K. Tokuda, H. Zen, A. W. Black. *An HMM-based speech synthesis system applied to English*, Proceedings of 2002 IEEE SSW, lk 227-230, 2002.

10. S. S. Stevens, J. Volkman, E. Newman. *A scale for the measurement of the psychological magnitude pitch*, Journal of the Acoustical Society of America, 8(3), lk 185-190, 1937.

11. T. Yoshimura, *Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for HMM-based text-to-speech systems*, PhD thesis, Nagoya Institute of Technology, 2002.

12. T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura. *Simultaneous Modeling of Spectrum, Pitch and Duration in HMM-Based Speech Synthesis*, Proceedings of EUROSPEECH, 5, lk 2347-2350, 1999.

13. A. Dempster, N. Laird, D. Rubin. *Maximum likelihood from incomplete data via the EM algorithm*, Journal of Royal Statistics Society, 39, lk 1-38, 1977.

14. T. Fukada, K. Tokuda, T. Kobayashi, S. Imai. *An adaptive algorithm for mel-cepstral analysis of speech*, Proceedings ICASSP-92, 1, lk 137-140, 1992.

15. J. J. Odell. *The Use of Context in Large Vocabulary Speech Recognition*, PhD thesis, Cambridge University, 1995.

16. H. Zen, K. Tokuda, A. W. Black. *Statistical parametric speech synthesis*, Speech Commun, 51(11), lk 1039-1064, 2009.

17. T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura. *Duration Modeling in HMM-based Speech Synthesis System*, Proceedings of ICSLP, 2, lk 29-32, 1998.

18. K. Tokuda, K. Oura, K. Hashimoto, S. Shiota, H. Zen, J. Yamagishi, T. Toda, T. Nose, S. Sako, A. W. Black. *HTS slides version 2.2 beta*, http://hts.sp.nitech.ac.jp/archives/2.2beta/HTS_Slides.zip [Viimati külastatud 10.05.2012]

19. H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, K. Tokuda. *The HMM-based Speech Synthesis System (HTS) Version 2.0*, Proceedings of SSW6-2007, lk 294-299, 2007.

20. K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura. *Speech parameter generation algorithms for HMM-based speech synthesis*, Proceedings of ICASSP 2000, 3, lk 1315-1318, 2000.

21. Y. Takamido, K. Tokuda, T. Kitamura, T. Masuko, T. Kobayashi. *A study of relation between speech quality and amount of training data in HMM-based TTS system*, ASJ Spring meeting, 2-10-14, lk 291-292, 2002 (jaapani keeles).

22. T. Masuko, K. Tokuda, T. Kobayashi, S. Imai. *Voice characteristics conversion for HMM-based speech synthesis system*, Proceedings of ICASSP-97, lk 1611-1614, 1997.

23. S. Imai. *Cepstral analysis synthesis on the mel frequency scale*, Proceedings of ICASSP-83, lk 93-96, 1983.

24. *Vocoder*, http://en.wikipedia.org/wiki/Vocoder [Viimati külastatud 10.05.2012]

25. H. Zen, T. Toda, M. Nakamura, T. Tokuda. *Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005*. IEICE Trans. Inf. Syst. E90-D, 1, lk 325-333, 2007.

26. M. Mihkla, A. Eek, E. Meister. *Creation of the Estonian Diphone Database for Text-to-Speech Synthesis*, Linguistica Uralica, 34(3), lk 334-340, 1998.

27. M. Mihkla, I. Kiissel, T. Nurk, L. Piits. *Eestikeelne korpuspõhine kõnesüntees*, http://www.keeletehnoloogia.ee/projektid/eestikeelne-korpusepohine-konesuntees [Viimati külastatud 10.05.2012]

28. M. Mihkla, L. Piits, T. Nurk, I. Kiissel. *Development of a Unit Selection TTS System for Estonian*, Proceedings of the Third Baltic Conference on Human Language Technologies, lk 181-187, 2008.

29. *Deutscher Wortschatz*, http://wortschatz.uni-leipzig.de/ [Viimati külastatud 10.05.2012]

30. T. Erelt, T. Leemets, S. Mäearu, M. Raadik. *Eesti Õigekeelsussõnaraamat*, Eesti Keele Sihtasutus, 2006.

31. S. King. *Speech synthesis without the right data*, Proceedings of SSW7-2010, 38, 2010.

32. *HMM-based Speech Synthesis System (HTS): Speaker dependent training demo for English*, http://hts.sp.nitech.ac.jp/archives/2.2/HTS-demo_CMU-ARCTIC-SLT.tar.bz2 [Viimati külastatud 10.05.2012]

33. *Hidden Markov Model Toolkit (HTK)*, http://htk.eng.cam.ac.uk/ [Viimati külastatud 10.05.2012]

34. M. Mihkla, I. Kiissel, T. Nurk, L. Piits. *Transcribing, structuring and temporal analysis of fluent speech corpus for unit selection TTS system for Estonian*, Computational Linguistics and Intellectual Technologies. Papers from Annual International Conference "Dialogue 2009“, lk 588-592, 2009.

35. A. Lundgren. *HMM-baserad talsyntes. An HMM-based Text-To-Speech System applied to Swedish*, Master thesis, Royal Institute of Technology, 2005 (rootsi keeles).

36. *Descàrrega de les veus Festcat*, http://www.talp.cat/festcat/download.php (katalaani keeles) [Viimati külastatud 10.05.2012]

37. *STRAIGHT, a speech analysis, modification and synthesis system*, http://www.wakayama-u.ac.jp/~kawahara/STRAIGHTadv/index_e.html
[Viimati külastatud 10.05.2012]

Lisad

Lisa 1. Süsteemis HTS kasutusel olev foneetiline tähestik häälemudeli ”Liisi” puhul

Foneem	Esitus süsteemis HTS	
	Lühike	Pikk
a	a	a:
e	e	e:
i	i	i:
o	o	o:
u	u	u:
õ	q	q:
ä	ae	ae:
ö	c	c:
ü	y	y:
j	j	j:
(w)	w	-
b	b	-
d	d	-
g	g	-
d'	D	-
p	p	p:
t	t	t
k	k	k
t'	T	T:
f	f	f:
s	s	s:
s'	S	S:
š	sh	sh:
l	l	l:
l'	L	L:
m	m	m:
n	n	n:
n'	N	N:
(ŋ)	ng	ng:
h	h	h:
v	v	v:
r	r	r:
(paus)	pau, SIL	-

Lisa 2. Süsteemis HTS kasutusel olevad foneemiklassid häälemudeli ”Liisi” puhul

```
QS "C-Vowel"
      {*-a+*,*-e+*,*-i+*,*-o+*,*-u+*,*-q+*,*-ae+*,*-c+*,*-y+*,*-a:+*,
*-e:+*,*-i:+*,*-o:+*,*-u:+*,*-q:+*,*-ae:+*,*-c:+*,*-y:+*}

QS "C-Semivowel"              {*-j+*,*-j:+*,*-w+*}

QS "C-Consonant"
      {*-b+*,*-d+*,*-g+*,*-D+*,*-p+*,*-t+*,*-k+*,*-T+*,,*-p:+*,*-t:+*,
*-k:+*,*-T:+*,*-f+*,*-s+*,*-S+*,*-sh+*,*-f:+*,*-s:+*,*-S:+*,*-sh:+*,
*-l+*,*-L+*,*-l:+*,*-L:+*,*-m+*,*-n+*,*-N+*,*-m:+*,*-n:+*,*-N:+*,*-ng+*,
*-ng:+*,*-h+*,*-v+*,*-r+*,*-h:+*,*-v:+*,*-r:+*}

QS "C-Stop"
      {*-b+*,*-d+*,*-g+*,*-D+*,*-p+*,*-t+*,*-k+*,*-T+*,*-p:+*,*-t:+*,
*-k:+*,*-T:+*}

QS "C-Nasal"
      {*-m+*,*-n+*,*-N+*,*-m:+*,*-n:+*,*-N:+*,*-ng+*,*-ng:+*}

QS "C-Fricative"
      {*-f+*,*-s+*,*-S+*,*-sh+*,*-f:+*,*-s:+*,*-S:+*,*-sh:+*,*-h+*,
*-h:+*,*-v+*,*-v:+*}

QS "C-Liquid"                 {*-l+*,*-L+*,*-l:+*,*-L:+*,*-r+*,*-r:+*}

QS "C-Coronal_Consonant"
      {*-n+*,*-t+*,*-d+*,*-s+*,*-sh+*,*-l+*,*-r+*,*-n:+*,*-t:+*,*-s:+*,
*-sh:+*,*-l:+*,*-r:+*}

QS "C-Non_Coronal"
      {*-b+*,*-p+*,*-k+*,*-T+*,*-D+*,*-g+*,*-p:+*,*-k:+*,*-T:+*,*-f+*,
*-S+*,*-f:+*,*-S:+*,*-L+*,*-L:+*,*-m+*,*-N+*,*-m:+*,*-N:+*,*-ng+*,
*-ng:+*,*-h+*,*-v+*,*-h:+*,*-v:+*}

QS "C-Sibilant_Consonant"     {*-s+*,*-S+*,*-sh+*,*-s:+*,*-S:+*,*-sh:+*}

QS "C-Front"
      {*-e+*,*-i+*,*-ae+*,*-c+*,*-y+*,*-e:+*,*-i:+*,*-ae:+*,*-c:+*,
*-y:+*,*-b+*,*-p+*,*-p:+*,*-m+*,*-m:+*,*-f+*,*-f:+*,*-v+*,*-v:+*}

QS "C-Central"
      {*-t+*,*-d+*,*-T+*,*-D+*,*-t:+*,*-T:+*,*-l+*,*-n+*,*-r+*,*-l:+*,
*-n:+*,*-r:+*,*-L+*,*-N+*,*-L:+*,*-N:+*,*-s+*,*-S+*,*-s:+*,*-S:+*,*-j+*,
*-j:+*,*-r+*,*-r:+*}

QS "C-Back"
      {*-a+*,*-o+*,*-u+*,*-q+*,*-a:+*,*-o:+*,*-u:+*,*-q:+*,*-w+*,*-k+*,
*-g+*,*-k:+*,*-h+*,*-h:+*,*-ng+*,*-ng:+*}

QS "C-Front_Vowel"
      {*-e+*,*-i+*,*-ae+*,*-c+*,*-y+*,*-e:+*,*-i:+*,*-ae:+*,*-c:+*,
*-y:+*}

QS "C-Back_Vowel"
      {*-a+*,*-o+*,*-u+*,*-q+*,*-a:+*,*-o:+*,*-u:+*,*-q:+*}

QS "C-Front_Consonant"
      {*-b+*,*-p+*,*-p:+*,*-m+*,*-m:+*,*-f+*,*-f:+*,*-v+*,*-v:+*}

QS "C-Central_Consonant"
      {*-t+*,*-d+*,*-T+*,*-D+*,*-t:+*,*-T:+*,*-l+*,*-n+*,*-r+*,*-l:+*,
*-n:+*,*-r:+*,*-L+*,*-N+*,*-L:+*,*-N:+*,*-s+*,*-S+*,*-s:+*,*-S:+*,*-r+*,
*-r:+*}

QS "C-Back_Consonant"
      {*-k+*,*-g+*,*-k:+*,*-h+*,*-h:+*,*-ng+*,*-ng:+*}

QS "C-Long"
      {*-a:+*,*-e:+*,*-i:+*,*-o:+*,*-u:+*,*-q:+*,*-ae:+*,*-c:+*,*-y:+*,
```

```
*-p:+*,*-t:+*,*-k:+*,*-T:+*,*-f:+*,*-s:+*,*-S:+*,*-l:+*,*-L:+*,*-m:+*,
*-n:+*,*-N:+*,*-ng:+*,*-h:+*,*-j:+*,*-v:+*,*-r:+*}

QS "C-Short"
      {*-a+*,*-e+*,*-i+*,*-o+*,*-u+*,*-q+*,*-ae+*,*-c+*,*-y+*,*-b+*,
*-p+*,*-t+*,*-d+*,*-k+*,*-g+*,*-T+*,*-D+*,*-f+*,*-s+*,*-S+*,*-l+*,*-L+*,
*-m+*,*-n+*,*-N+*,*-ng+*,*-h+*,*-j+*,*-w+*,*-v+*,*-r+*}

QS "C-Long_Vowel"
      {*-a:+*,*-e:+*,*-i:+*,*-o:+*,*-u:+*,*-q:+*,*-ae:+*,*-c:+*,*-y:+*}

QS "C-Short_Vowel"
      {*-a+*,*-e+*,*-i+*,*-o+*,*-u+*,*-q+*,*-ae+*,*-c+*,*-y+*}

QS "C-High_Vowel"             {*-i+*,*-i:+*,*-u+*,*-u:+*,*-y+*,*-y:+*}

QS "C-Medium_Vowel"
      {*-e+*,*-o+*,*-c+*,*-q+*,*-e:+*,*-o:+*,*-c:+*,*-q:+*}

QS "C-Low_Vowel"              {*-a+*,*-a:+*,*-ae+*,*-ae:+*}

QS "C-Rounded_Vowel"
      {*-o+*,*-u+*,*-c+*,*-y+*,*-o:+*,*-u:+*,*-c:+*,*-y:+*}

QS "C-Unrounded_Vowel"
      {*-a:+*,*-e:+*,*-i:+*,*-q:+*,*-ae:+*,*-a+*,*-e+*,*-i+*,*-q+*,
*-ae+*}

QS "C-IVowel"                 {*-i:+*,*-i+*,*-y+*,*-y:+*}

QS "C-EVowel"                 {*-e:+*,*-q:+*,*-c:+*,*-e+*,*-q+*,*-c+*}

QS "C-AVowel"                 {*-a:+*,*-ae:+*,*-a+*,*-ae+*}

QS "C-OVowel"                 {*-o:+*,*-o+*}

QS "C-UVowel"                 {*-u:+*,*-u+*}

QS "C-Voiced"
      {*-l+*,*-L+*,*-l:+*,*-L:+*,*-m+*,*-n+*,*-N+*,*-ng+*,*-m:+*,*-n:+*,
*-N:+*,*-ng:+*,*-j+*,*-w+*,*-v+*,*-r+*,*-j:+*,*-v:+*,*-r:+*,*-a+*,*-e+*,
*-i+*,*-o+*,*-u+*,*-q+*,*-ae+*,*-c+*,*-y+*,*-a:+*,*-e:+*,*-i:+*,*-o:+*,
*-u:+*,*-q:+*,*-ae:+*,*-c:+*,*-y:+*}

QS "C-Unvoiced_Consonant"
      {*-b+*,*-p+*,*-t+*,*-d+*,*-k+*,*-g+*,*-T+*,*-D+*,*-p:+*,*-t:+*,
*-k:+*,*-T:+*,*-f+*,*-s+*,*-S+*,*-f:+*,*-s:+*,*-S:+*,*-h+*,*-h:+*}

QS "C-Voiced_Consonant"
      {*-l+*,*-L+*,*-l:+*,*-L:+*,*-m+*,*-n+*,*-N+*,*-ng+*,*-m:+*,*-n:+*,
*-N:+*,*-ng:+*,*-v+*,*-r+*,*-v:+*,*-r:+*}

QS "C-Front_Stop"             {*-b+*,*-p+*,*-p:+*}

QS "C-Central_Stop"           {*-T+*,*-D+*,*-t+*,*-d+*,*-T:+*,*-t:+*}

QS "C-Back_Stop"              {*-k:+*,*-k+*,*-g+*}

QS "C-Voiced_Fricative"       {*-v+*,*-v:+*}

QS "C-Unvoiced_Fricative"
      {*-f+*,*-s+*,*-S+*,*-sh+*,*-f:+*,*-s:+*,*-S:+*,*-sh:+*,*-h+*,
*-h:+*}

QS "C-Front_Fricative"        {*-f+*,*-v+*,*-f:+*,*-v:+*}

QS "C-Central_Fricative"      {*-s+*,*-S+*,*-sh+*,*-s:+*,*-S:+*,*-sh:+*}

QS "C-Back_Fricative"         {*-h+*,*-h:+*}

QS "C-Silences"               {*-pau+*,*-SIL+*}
```

Lisa 3. Treenitud häälemudelitega genereeritud näitelaused

Tööga kaasas olev CD-plaat sisaldab järgnevat:

- kaustades `liisi_100/` ja `liisi_500/` on vastavalt 100- ja 500-lauselisel korpusel treenitud häälemudeliga genereeritud sünteeskõne näitelaused wav-formaadis ning teksti kujul;
- kaustades `näitelause_1/ .. näitelause_8/` on igas üks lause genereeritud erinevatel korpustel (”Tõnu” alamkorpus, Liisi alamkorpused mahuga 500, 1000 ja 2000 lauset ning Liisi alamkorpused mahuga 250, 487 ja 2000 lauset, kasutades foneetikakirjelduses vaid kvalitatiivfoneeme) treenitud häälemudelitega ja antud lause teksti kujul.