HMM-POHJAISEN PUHESYNTEESIN LAADUN PARANTAMINEN GLOTTISPULSSIKIRJASTON AVULLA

Tuomo Raitio[1], **Antti Suni**[2], **Hannu Pulakka**[1], **Martti Vainio**[2], **Paavo Alku**[1]

[1] Aalto-yliopisto, Signaalinkäsittelyn ja akustiikan laitos
PL 13000/Otakaari 5 A, 02150 Espoo
etunimi.sukunimi@tkk.fi

[2] Helsingin yliopisto, Puhetieteiden laitos
PL 9/Siltavuorenpenger 5 A, 00014 Helsingin yliopisto
etunimi.sukunimi@helsinki.fi

1 JOHDANTO

Markovin piilomalleihin (hidden Markov Model, HMM) perustuvien parametristen puhesyntetisaattorien [1] käyttö on viime aikoina yleistynyt niiden monipuolisten muokkausmahdollisuuksien ja pienen tilantarpeen takia, mutta niiden tuottaman puheäänen laatu on pysynyt huonompana hyvälaatuisiin konkatenaatiosynteeseihin verrattuna. Huonompi laatu johtuu pääasiassa kolmesta eri syystä: liian yksinkertaisista vokoderitekniikoista, akustisten mallien epätarkkuudesta ja parametrien liiallisesta keskiarvoistumisesta [1]. Tässä työssä pyritään parantamaan HMM-pohjaisen synteesin laatua ja luonnollisuutta paremman vokoderitekniikan avulla.

Useimmat HMM-pohjaiset syntetisaattorit perustuvat tavalla tai toisella puheentuoton lähde-suodin malliin [2], jossa ääni syntyy kurkunpäässä sijaitsevasta glottiksesta lähtevästä herätteestä, joka tämän jälkeen suodattuu ääntöväylän akustisella suotimella. Puhesynteesin huono laatu johtuu pitkälti juuri riittämättömistä herätemalleista. Viime aikoina ääniherätteen mallinnusta onkin tutkittu runsaasti paremman äänenlaadun toivossa. Perinteisen impulssiherätteen korvaajiksi on ehdotettu esimerkiksi impulssin ja kohinan yhteisherätettä (mixed-excitation) [3] tai glottisilmavirtauksen differentioitua aaltomuotomallia, niin kutsuttua Liljencrants-Fant (LF) -mallia [4]. Vaikka nämä metodit ovat vähentäneet puheen häiritsevää konemaisuutta, puheen luonnollisuuden ja puhujan äänen eri ominaisuuksien mallintamisessa on vielä huomattavasti parantamisen varaa. Koska äänilähteen tarkka mallintaminen on osoittautunut erittäin vaikeaksi, viime aikoina mallien sijasta onkin alettu käyttää itse aidosta puheesta estimoituja glottispulsseja, mikä on selvästi parantanut synteettisen äänen luonnollisuutta [5].

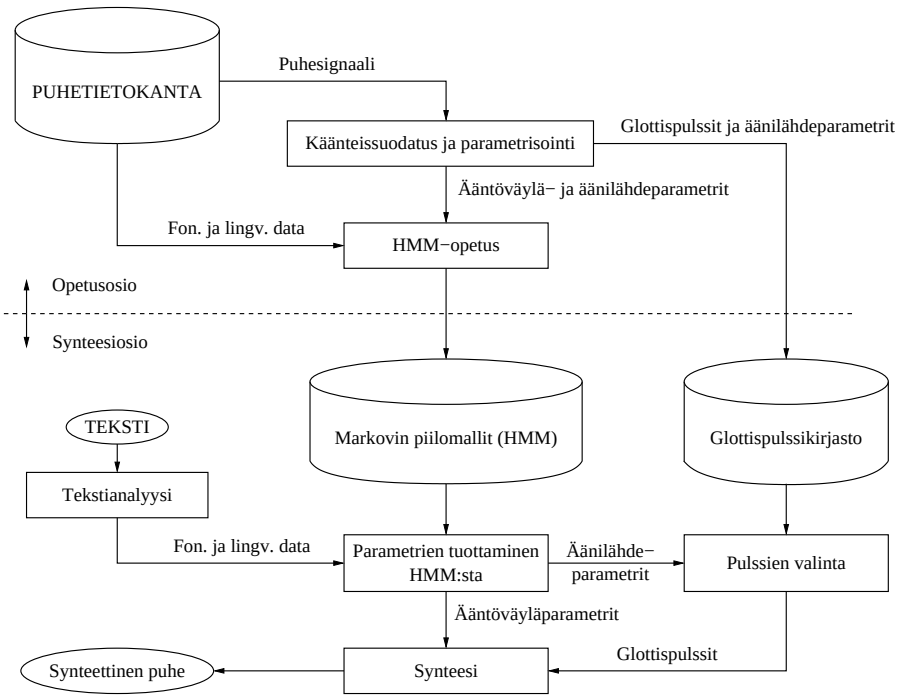
Aalto-yliopiston Signaalinkäsittelyn ja akustiikan laitoksen sekä Helsingin yliopiston Puhetieteiden laitoksen yhdessä kehittämä HMM-pohjainen puhesynteesijärjestelmä [6] perustuu lähde-suodin malliin, jossa heräte luodaan äänilähteen käänteissuodatuksella [7] puheesta estimoidusta aidosta glottispulssista. Tällä tekniikalla tuotetun puheen laatu on parempaa verrattuna tavallisin herätemallein tuotettuun puheeseen [6, 8], mutta yhden pulssin käyttäminen rajoittaa puheen eri ominaisuuksien, mm. puhujan identiteetin ja eri ääntötapojen mallintamista. Tässä työssä kyseistä synteesimenetelmää on laajennettu sisältämään satoja tai jopa tuhansia glottispulsseja sisältävä kirjasto, josta valitaan kulloiseenkin äänteeseen, kontekstiin ja puhujaan sopivat herätepulssit.

2 PUHESYNTEESIJÄRJESTELMÄ

2.1 Yleiskatsaus

Työssä esitetyn parametrisen HMM-pohjaisen puhesynteesijärjestelmän tavoitteena on tuottaa hyvälaatuista ja luonnollista synteettistä puhetta eri puhetyyleillä ja puhujilla. Syntetisaattorissa hyödynnetään äänilähteen käänteissuodatusta [7], jonka avulla puheesta erotetaan toisistaan puheentuoton kaksi eri komponenttia: glottiksessa syntyvä äänilähde ja ääntöväylän suodin. Tällä tavoin äänilähde voidaan erikseen parametrisoida ja siitä voidaan rakentaa glottispulssikirjasto synteesiä varten.

Kuvassa 1 on esitetty puhesynteesijärjestelmän rakenne. Järjestelmä koostuu kahdesta eri osiosta: opetus- ja synteesiosiosta. Opetusvaiheessa tallennetun äänitietokannan puhesignaalit jaetaan äänilähteeseen ja ääntöväyläsuotimeen käänteissuodatuksen avulla. Tämän jälkeen äänilähde ja ääntöväyläsuodin parametrisoidaan ja äänilähteestä erotetaan yksittäiset glottispulssit. Äänilähteen parametridatasta ja glottispulsseista rakennetaan kirjasto, jota käytetään synteesivaiheessa. Tämän jälkeen puheesta saadut parametrit ja puheen foneettinen sekä lingvistinen informaatio opetetaan Markovin piilomalleihin (HMM), jotka kuvaavat tilastollisesti puheäänen parametrien ja foneettisten sekä lingvististen kontekstien suhdetta. Itse parametrit mallinnetaan Gaussian mixture -mallien (GMM) avulla. Synteesivaiheessa HMM:sta tuotetaan foneettisia ja lingvistisiä konteksteja vastaavat puheäänen parametrit syntetisoitavan tekstin mukaan. Äänilähteen luomista varten glottispulssikirjastosta valitaan äänilähteen parametreja parhaiten vastaavat pulssit. Vokooderi luo herätteen valituista pulsseista ja suodattaa herätteen ääntöväylän suotimella, jolloin saadaan aikaan synteettistä puhetta.



Kuva 1: *Puhesynteesijärjestelmän rakenne.*

2.2 Parametrisointi

Kuvassa 2 on esitetty puheen parametrisoinnin vuokaavio. Ennen parametrisointia, signaali ylipäästösuodatetaan, jotta signaalista saadaan pois pienitaajuiset puheeseen kuulumattomat taajuuskomponentit, ja ikkunoidaan 5 ms välein kahdenlaisiin kehyksiin: 25 ms kehyksiin puheen spektrille sekä energialle ja 44 ms kehyksiin äänilähdeparametreille ja glottispulssien erottamiseen.

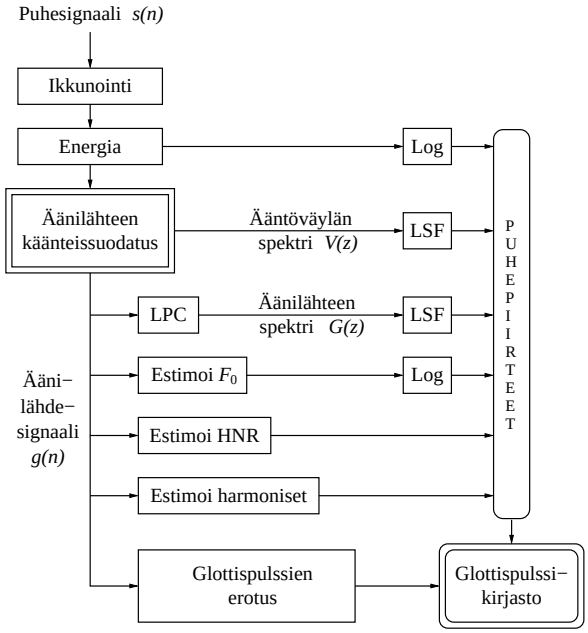
Molemmissa kehyksissä puhesignaali erotetaan äänilähteeseen ja ääntöväylän parametreihin käyttäen automaattista iteratiivista äänilähteen käänteissuodatusta (IAIF) [7]. IAIF iteratiivisesti kumoaa ääntöväylän ja huulisäteilyn vaikutuksen puhesignaalista käyttäen all-pole mallinnusta, minkä tuloksena saadaan estimoitua äänilähdesignaali sekä ääntöväylän spektri. Jotta äänilähteen vaihtelu (mm. spektrin kaltevuus) eri äänteistä ja puhetyyleistä johtuen saadaan mallinnettua, äänilähdesignaalista estimoidaan lineaariprediktiota (LPC) käyttäen äänilähteen spektri. Ääntöväylän ja äänilähteen spektrit muutetaan paremmin tilastolliseen mallinnukseen sopivaan muotoon viivaspektritaajuuksisksi (LSF). Soinnittomissa kehyksissä puheen spektriä mallinnetaan tavallisen LPC:n avulla.

Puheen perustaajuus (F_0) estimoidaan autokorrelaatiomenetelmällä äänilähdesignaalista. Puheen soinnillisuus estimoidaan äänilähdesignaalista mittaamalla lähteen spektristä harmonisten piikkien ja näiden välisen kohinan magnitudisuhde viideltä eri ERB-taajuuskaistalta. Näiden lisäksi lasketaan kymmenen ensimmäisen harmonisen piikin magnitudit kuvaamaan tarkasti pienien taajuuksien spektrin kaltevuutta (spectral tilt).

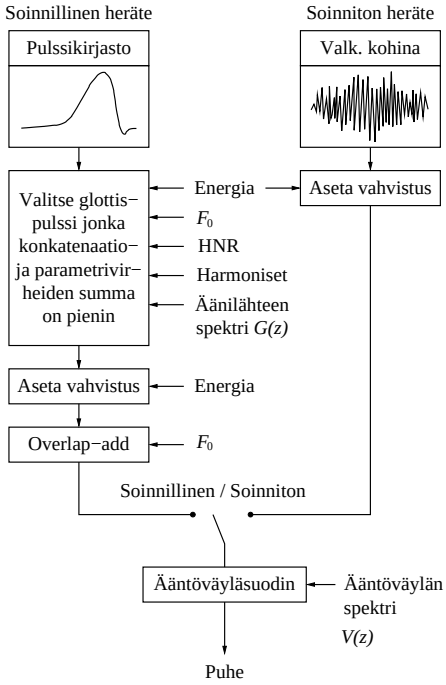
Glottispulssikirjastoa varten äänilähteestä erotetaan yksittäiset glottispulssit. Ensin äänilähdesignaalista estimoidaan glottiksen sulkeutumisajankohdat (glottal closure instant, GCI) etsimällä signaalista minimiarvoa perusjaksojen välein. Tämän jälkeen jokainen kokonainen kahden perusjakson mittainen glottispulssipari erotetaan signaalista siten, että GCI jää kehyksen keskikohtaan. Glottispulssikehys ikkunoidaan Hann-ikkunalla ja sen energia normalisoidaan, jonka jälkeen pulssi tallennetaan kirjastoon pulssia kuvaavien äänilähdeparametrien kanssa. Näiden lisäksi kirjastoon tallennetaan myös 10 ms pituinen alinäytteistetty pulssi kuvaamaan pulssin aaltomuotoa.

2.3 Synteesi

Kuvassa 3 on esitetty puheen syntetisoinnin vuokaavio. Puheen herätesignaali koostuu sekä soinnillisesta että soinnittomasta äänilähteestä. Soinnillinen äänilähde luodaan glottispulssikirjastosta valituista pulsseista. Sopivin pulssi kullekin aikaindeksille valitaan minimoimalla yhteisvirhe, joka koostuu parametrivirheestä ja konkatenaatiovirheestä. Parametrivirhe on HMM:ien tuottamien äänilähdeparametrien ja kunkin glottispulssin äänilähdeparametrien erotus, ja se kuvaa glottispulssin sopivuutta syntetisoitavaan äänteeseen ja kontekstiin. Parametrivirheen minimoinnilla pyritään siihen, että äänilähteen ominaisuuksiltaan (perustaajuus, spektrin kaltevuus, kohinan määrä) sopiva pulssi tulee valituksi. Koska perustaajuus sisältyy optimoitaviin parametreihin, oikean perustaajuuden omaava pulssi tulee automaattisesti valituksi, eikä pulsseja tarvitse erikseen muokata, toisin kuin yhden pulssin tekniikassa. Vain pulssin energia asetetaan HMM:sta saadun arvon mukaiseksi. Konkatenaatiovirhe on peräkkäisten glottispulssien



Kuva 2: *Puhesignaalin parametrisointi.*



Kuva 3: *Syntetisointi.*

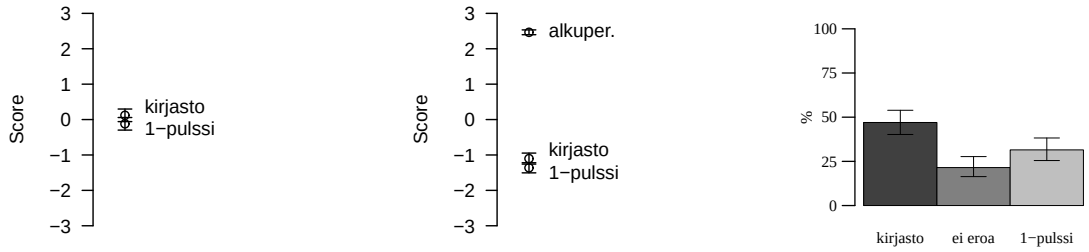
alinäytteistettyjen aaltomuotojen erotuksen neliöllinen keskiarvo, ja se kuvaa glottispulssien liitoskohdan sopivuutta. Konkatenaatiovirheen minimoinnilla pyritään siihen, että peräkkäiset pulssit eivät eroa liikaa toisistaan, mikä voi tuottaa epäjatkumia äänilähteeseen ja johtaa epätasaiseen äänenlaatuun.

Kussakin yhtäjaksoisessa soinnillisessa äänteessä pulssien valinta optimoidaan yhteisvirheen perusteella käyttämällä Viterbi-algoritmia. Valintaprosessia voidaan säätää muuttamalla äänilähteen parametreille, parametrivirheelle ja konkatenaatiovirheelle asetettuja erillisiä painokertoimia. Pulssien valinnan jälkeen pulssit liitetään toisiinsa overlap-add tekniikalla perustaajuuden mukaan.

Soinniton heräte koostuu valkoisesta kohinasta, jonka vahvistus määräytyy HMM:sta saadun arvon perusteella. Soinnillinen ja soinniton heräte yhdistetään yhdeksi herätteeksi perustaajuusinformaation perusteella. Ennen suodatusta, ääntöväylän suotimen parametreja muokataan [9], millä kompensoidaan tilastollisen mallinnuksen keskiarvoistumista. Näin suotimen formantit saadaan voimakkaammiksi ja siten luonnollisemmiksi. Lopulta äänilähde suodatetaan ääntöväylän suotimella, jolloin saadaan aikaan puhetta. Ääniesimerkkejä on saatavilla osoitteessa www.helsinki.fi/speechsciences/synthesis/samples.html.

3 KUUNTELUKOKEET

Uuden puhesynteesijärjestelmän laatua tutkittiin järjestämällä kuuntelukokeita. Ensiksi, kaksi eri puhesynteesijärjestelmää, yhden pulssin tekniikkaa hyödyntävä järjestelmä sekä glottispulssikirjastoa hyödyntävä järjestelmä, opetettiin 600 lauseen suomenkielisen



(a) (b) (c)

Kuva 4: *CCR testin tulokset a) mv ja b) rjs -tietokannoilla, sekä c) samankaltaisuustestin tulokset rjs-tietokannalla: yhden pulssin tekniikka (1-pulssi), glottispulssikirjastotekniikka (kirjasto) ja luonnollinen puhe (alkuper.). 95% luotettavuusvälit on ilmoitettu kullekin tulokselle.*

miehen (*mv*) puhumalla puhetietokannalla. Tietokannan 30:stä ensimmäisestä lauseesta luotiin pulssikirjasto, joka sisälsi 22044 glottispulssia. Comparison category rating (CCR) testiä käytettiin puheen laadun arvioinnissa. CCR-testissä testihenkilö kuulee ääninäyteparin, ja koehenkilön tehtävänä on arvioida comparison mean opinion score (CMOS) -asteikolla ääninäyteparin laatuero. Molemmilla järjestelmillä tuotettiin kymmenen synteettistä lausetta. Kymmenen suomenkielistä koehenkilöä arvioi yhteensä 30 ääninäyteparia. Metodien paremmuus laskettiin keskiarvoistamalla kummallekin metodille annetut arviot. Testin tulokset on esitetty Kuvassa 6a.

Seuraavaksi molemmat järjestelmät opetettiin englanninkielisen miespuolisen ammattipuhujan puhetietokannalla (*rjs*). Kolmestakymmenestä ensimmäisestä lauseesta luotiin jälleen pulssikirjasto, joka sisälsi 23332 glottispulssia. CCR-testiä käytettiin jälleen arvioimaan puheen laatua, mutta tässä testissä arviointiin otettiin mukaan myös alkuperäiset lauseet. Kymmenen suomenkielistä mutta englantia hyvin puhuvaa koehenkilöä arvioi yhteensä 70 ääninäyteparia. Testin tulokset on esitetty Kuvassa 6b.

Lopulta koehenkilöitä pyydettiin arvioimaan synteettisen puheen (*rjs*-puhetietokanta) samankaltaisuutta verrattuna alkuperäiseen puheeseen. Testissä koehenkilöille esitettiin kolme näytettä, *A*, *B* ja *Ref*, ja tehtävänä oli valita, kumpi vaihtoehdoista *A* vai *B* kuulosti enemmän referenssipuhujalta *Ref*. Koehenkilöillä oli myös mahdollisuus valita, että kumpikaan vaihtoehdoista ei ole samankaltaisempi. Samankaltaisuustestin tulokset on esitetty Kuvassa 6c.

Molemmissa CCR-testeissä pulssikirjastoa hyödyntävällä tekniikalla tuotettu synteettinen puhe on keskimäärin arvioitu paremmaksi kuin yhden pulssin tekniikalla tuotettu puhe. Tulokset eivät kuitenkaan ole tilastollisesti merkittäviä ristiriitaisten tulosten takia; jotkin kuulijat pitivät yhden pulssin tekniikan tuottamaa tasaista laatua parempana. Puheen samankaltaisuudessa pulssikirjastotekniikka on tilastollisesti parempi.

4 KESKUSTELU

Kuuntelukokeiden tulokset osoittavat, että pulssikirjastoa hyödyntävä tekniikka tuottaa yhtä hyvää tai parempilaatuista puhetta kuin yhden pulssin tekniikka. Koska yhden

pulssin tekniikan tuottama puheen laadun on jo todettu olevan hyvin korkea [6], myös kyseessä olevan pulssikirjastotekniikan voidaan sanoa olevan selvästi tavallisia herätemalleja parempi.

Tässä työssä esitetyn herätemallin edut johtuvat pulssikirjastoon tallentuvasta luonnollisesta äänilähteestä. Erityisesti osittain soinnilliset äänteet, kuten [v,f,h,z] toistuvat huomattavasti luonnollisemmin kyseisellä tekniikalla, koska yhden pulssin tekniikka ei kykene riittävästi muokkaamaan pulssia saadakseen aikaan luonnollista muistuttavaa herätettä. Kuuntelukokeet kuitenkin osoittivat, että esitetty metodi jakoi kuuntelijoiden mielipiteitä: vaikka monipuolisempi heräte koetaan yleisesti luonnollisempana, jotkin kuulijat pitivät herätettä vähemmän luonnollisena pienten epätasaisuuksien takia. Esitetyn herätemallin kehitystyö jatkuu jotta herätteestä saataisiin poistettua epäluonnolliset epätasaisuudet.

KIITOKSET

Tutkimusta on tukenut Suomen Akatemia ja MIDE UI-ART.

VIITTEET

[1] BLACK A, ZEN H, & TOKUDA K, Statistical parametric speech synthesis, volume 4, pages 1229–1232, Apr. 2007.

[2] FANT G, *Acoustic Theory of Speech Production*, Mouton, The Hague, 1960.

[3] YOSHIMURA T, TOKUDA K, MASUKO T, KOBAYASHI T, & KITAMURA T, Mixed excitation for HMM-based speech synthesis, in *Proc. Eurospeech*, pages 2259–2262, 2001.

[4] FANT G, LILJENCRANTS J, & LIN Q, A four-parameter model of glottal flow, *STL-QPSR*, **4**(1985), 1–13.

[5] DRUGMAN T, WILFART G, MOINET A, & DUTOIT T, Using a pitch-synchronous residual codebook for hybrid HMM/frame selection speech synthesis, in *Proc. ICASSP*, pages 3793–3796, Apr. 2009.

[6] RAITIO T, SUNI A, YAMAGISHI J, PULAKKA H, NURMINEN J, VAINIO M, & ALKU P, HMM-based speech synthesis utilizing glottal inverse filtering, *IEEE Trans. on Audio, Speech, and Lang. Proc.*, **19**(2011) 1, 153–165.

[7] ALKU P, Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering, *Speech Commun.*, **11**(1992) 2-3, 109–118.

[8] SUNI A, RAITIO T, VAINIO M, & ALKU P, The GlottHMM speech synthesis entry for Blizzard Challenge 2010, in *The Blizzard Challenge 2010 workshop*, 2010, http://festvox.org/blizzard.

[9] RAITIO T, SUNI A, PULAKKA H, VAINIO M, & ALKU P, Comparison of formant enhancement methods for HMM-based speech synthesis, in *SSW7*, pages 334–339, Sep. 2010.