8 Puhesynteesi

Puhesynteesillä voidaan periaatteessa ymmärtää mitä hyvänsä puheen syntetisointia, eli esimerkiksi puhekoodekin suorittamaa puhesignaalin generointia vastaanotettujen parametrien perusteella, tai tietokoneella ajettua proseduuria, joka laskee eräänlaisen puhesignaalin esityksen annetusta tekstistä. Koska koodekit ovat varsinaisesti Puheen koodaus -kurssin asiaa, tässä kappaleessa keskitytään nimenomaan *tekstistä puheeksi* -synteesiin (*text-to-speech*, TTS), johon viitataankin jatkossa yksinkertaisesti nimellä puhesynteesi. On kuitenkin hyvä pitää mielessä, että synteesimuodosta riippumatta puhesignaalin laadulle asetetaan samanlaisia tavoitteita. Tähän aiheeseen palataan lyhyen TTS-motivoinnin jälkeen, ja kappaleen loppuosa omistetaan erilaisille TTS:n toteuttamiseen liittyville asioille.

Tekstistä puheeksi -synteesi on laaja tutkimusala, johon on jo joidenkin vuosikymmenien ajan panostettu runsaasti voimavaroja. Päteviä syitä löytyy useita. Eräs mielenkiintoisimmista (joskin samalla melko pitkälle tulevaisuuteen tähyilevä) seikka on se, että toimiva TTS-systeemi, yhdistettynä toimivaan puheentunnistimeen, on itse asiassa erittäin tehokas puheenkoodausmenetelmä (Huang, Acero, Hon, 2001). Se tarjoaa ylivertaisen pakkaussuhteen ja joustavat mahdollisuudet valita syntetisoidun puheen tyypin (esim. hengästynyt tai kähisevä), perustaajuuden vaihtelualueineen, puheen rytmin sekä paljon muita efektejä. Lisäksi viestin sisällön muuttaminen käy paljon kätevämmin tekstiä vaihtamalla kuin äänittämällä signaali uudelleen. Valitettavasti tällaista järjestelmää ei kuitenkaan laajalle sanavarastolle ole vielä olemassa.

Puhesynteesille löytyy toki myös lähempänä toteutumista olevia tärkeitä sovelluksia, kuten erilaisia puhelinpalveluita, joissa useinkin päivittyvät tiedot voidaan syntetisaattorin avustuksella välittää kysyjälle. Suuri apu puhesyntetisaattoreista on tietysti näkövammaisille ja puhekykynsä menettäneille henkilöille. Muitakin arkisia sovelluksia on helppo keksiä: viestien ja uutisten kuunteleminen lukemisen sijaan, autoissa käytettävät hands-free-toiminnot ja niin edelleen.

8.1 Syntetisoidun puhesignaalin laatu

Tavallisimpia puheen laadun kriteereitä ovat ymmärrettävyys, luonnollisuus ja miellyttävyys. Nämä ovat monitahoisia asioita, jotka riippuvat toisistaan, joten kokonaisvaltaisesti laadukkaan puhesignaalin muodostavat lukuisat eri tekijät yhdessä. Paitsi että taustakohina, musikaalinen kohina, mumina ja erilaiset rapsahdukset ja poksahdukset on saatava eliminoitua, täytyisi puheesta saada myös tehtyä persoonallisen kuuloista ja vivahteikasta. Puhujan tunnetilan välittyminen puheen kautta olisi myös suotavaa, sillä se tuo luonnollisuutta ja elävyyttä. Sinänsä puheen laadusta erillinen, mutta syntetisaattorien tapauksessa kiinnostava seikka on myös se, kuinka riippuvainen syntetisaattori on käytetystä ohjelmistoalustasta tai ympäristöstä ja kuinka hyvin se toimii siinä sovelluksessa, johon se on tarkoitettu. Tietokoneiden laskentatehon alituisesti kasvaessa yhä monimutkaisempia syntetisaattoreita voidaan käyttää ihan "kotikoneellakin".

Kuten jo puheen ehostuksen yhteydessä todettiin, puheen laadun mittaaminen on enemmän tai vähemmän epämääräistä puuhaa, eikä voida nimetä mitään menetelmää, joka antaisi absoluuttisesti oikeita ja kaiken kattavia tuloksia. Erilaisia testejä yhdistämällä saadaan kuitenkin usein sopiva määrä tietoa siitä, kuinka laadukasta puhetta tietty järjestelmä tuottaa. Toki testien suuren skaalan ja monien epävarmuustekijöiden vuoksi eri tahojen julkaisemia tuloksia voi olla hankala vertailla keskenään. Lisäksi monien eri testien, varsinkin kuuntelutestien, järjestäminen on hyvin työlästä ja kallista.

Jos nyt kustannusseikat ja muut hankaluudet jätetään kuitenkin huomiotta, on puhesynteesin lop-

putulosta luonnollisesti viisainta tarkastella monelta eri taholta, jotta voidaan vetää johtopäätöksiä eri moduulien, kuten lingvistisen tai prosodisen analyysin toimivuudesta (näistä lisää hetken kuluttua). Puheen ymmärrettävyyttä voidaan arvioida esim. testaamalla, kuinka hyvin kuuntelijat erottavat eri konsonantteja toisistaan syntetisoidussa puheessa. Konsonantteja onkin vaikeampaa tuottaa puhesynteesillä kuin vokaaleja. Kontekstiriippuvuuden poistamiseksi käytetään usein hyvin lyhyitä testiilmauksia, kuten tavuja, tai sitten sanoja jotka eivät tarkoita mitään. Pidemmän aikavälin ymmärrettävyyttä testattaessa käytetään puolestaan kokonaisia lauseita, jolloin virheet tietyissä äänteissä eivät välttämättä häiritse viestin sisällön ymmärtämistä. Syntetisaattorin onnistumista prosodisten piirteiden välittämisessä onkin jo melko epämääräistä testata, mutta tähänkin voidaan käyttää erilaisia kuuntelijoiden mielipiteitä tunnustelevia testejä. Yleistä äänenlaatua, eli vääristymien, muminan, rahinan ja muiden efektien määrää mittaavia kuuntelutestejä on sen sijaan kehitetty jo pidempään ja niiden käytössä esiintyy tiettyä säännönmukaisuutta. Testimenetelmiä ei tässä käydä läpi, mutta tietoa tavallisimmista objektiivisista mittareista ja kuuntelutesteistä löytää lähes mistä tahansa puheenkäsittelyn kirjasta.

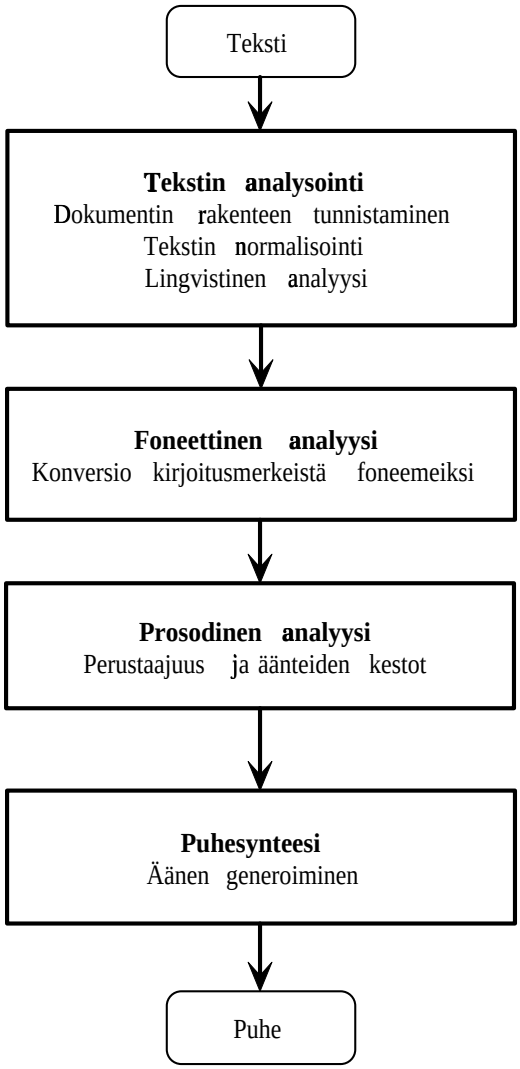
Nykypäivän puhesyntetisaattoreissa on vielä paljon puutteita puheen luonnollisuuden ja persoonallisuuden suhteeen. Ymmärrettävyys on kuitenkin jo hyvällä tasolla, joten monissa sovelluksissa puhesyntetisaattoreita voidaan jo oikeasti käyttää. Ymmärrettävyys vielä paranee, jos lisäinformaationa esitetään kasvoanimaatioita. Tämä ns. *audiovisuaalinen puhesynteesi* onkin uusimpia trendejä synteesin alalla (Lemmetty, 1999).

8.2 Puhesynteesin toteutuksesta

Tekstin muuttaminen puheeksi on karkeasti ottaen kaksivaiheinen prosessi: ensin teksti analysoidaan ja sitten generoidaan puhesignaali. Näistä ensimmäinen vaihe käsittää oheisessa lohkokaaviossa (kuva 1) itse tekstin analysoinnin lisäksi foneettisen analyysin, jossa kirjoitusmerkit tulkitaan foneemeiksi. Myös puhesignaalin generointi voidaan edelleen jakaa kahteen tehtävään: osasten etsintään tietokannasta, tai niiden generointiin, ja prosodisten piirteiden toteuttamiseen. Katsotaanpa seuraavaksi näitä vaiheita hieman tarkemmin.

Tekstin analysoinnissa on kysymys tekstin muuttamisesta sellaiseen muotoon, että siitä tulee "puhumiskelpoista". Yksinkertaisimmillaankin tässä lohkossa tehdään vähintään tekstin normalisointi, jossa numerot muutetaan lukusanoiksi, lyhenteet kirjoitetaan auki jne. Tässä käytetään tyypillisesti suurta joukkoa sääntöjä, jotka yrittävät huomioida mm. kielestä ja asiayhteydestä riippuvia tekijöitä. Haasteellisimpana tehtävänä tekstianalyysilohkossa on lingvistinen eli kielitieteellinen analyysi, joka tarkoittaa syntaktista ja semanttista, tekstin sisällön ymmärtämiseen tähtäävää analyysia. Tietokone ei tietenkään ihmisen lailla kykene sisältöä ymmärtämään, mutta tilastollisiin menetelmiin perustuen yritetään silti löytää todennäköisin vaihtoehto ilmaisun sisällölle, koska ääntäminen riippuu tietyissä tapauksissa sanan merkityksestä ja erityisesti kontekstista (esim. englannin kielen `record` substantiivina tai verbinä). Esimerkkinä kontekstiriippuvuudesta suomen kielessä olkoon vaikkapa konsonanttien kahdentumiset, kuten edellä olleessa ilmaisussa "yritetään löytää-(t)-todennäköisin...". Tekstin analysoinnin on tarkoitus myös tuottaa tietoa ilmausten prosodiasta, eli mm. erotella kysymys- ja toteamuslauseet toisistaan, jotta intonaatio voitaisiin sovittaa lausetyypin mukaan, ja tunnistaa taukojen paikkoja puheessa välimerkkien perusteella.

Foneettinen analyysi konvertoi kirjoitusjärjestelmän merkit ääntämisen mukaisiksi merkeiksi käyttäen jotakin foneettista aakkostoa. Aiemmin tällä kurssilla onkin jo tavattu IPA:n foneettinen aakkos-



Kuvio 1: *Lohkokaavio tekstistä puheeksi synteesistä. Alkuperäinen englanninkielinen kuva löytyy seuraavasta kirjasta sivulta 6: X. Huang, A. Acero, H.-W. Hon, Spoken Language Processing, Prentice Hall PTR, 2001.*

to, josta löytyvät paitsi foneemisymbolit, myös ääntämiseen ja aksentteihin liittyviä symboleja. Koska IPA:n symbolit ovat monimutkaisia ja sisältävät monia merkkejä, joita ei löydy tietokoneen näppäimistöltä, on IPA-järjestelmän rinnalle kehitetty myös paremmin tietokoneiden kanssa yhteensopivia, ASCII-merkkeihin perustuvia järjestelmiä, kuten SAMPA (*Speech Assessment Methods - Phonetic Alphabet*), *Worldbet* ja *Arpabet*. Yleisesti hyväksyttyä yhteistä foneettista aakkostoa ei kuitenkaan ole, ja tästä syystä monet puhesyntetisaattorit käyttävät todellisuudessa omia erityisiä foneettisia aakkostojaan (Lemmetty, 1999). Foneettisen analyysin haasteellisuus riippuu vahvasti kielestä - suomen kieli onkin tässä suhteessa helpoimmasta päästä, koska ääntämis- ja kirjoitusasu eivät kovin paljon eroa toisistaan.

Prosodiaan kuuluvat puheen rytmi, painotukset ja intonaatio (eli ns. puheen sävelkulku), joiden ominaisuuksia analysoidaan luonnollisesta puheesta ja tämän perusteella luodaan sääntöjä vastaavien ominaisuuksien tuottamiseksi synteettiseen puheeseen. Prosodialla on erittäin suuri merkitys puheen

ymmärrettävyyden kannalta ja lisäksi puheen prosodiset piirteet välittävät tietoa puhujan ominaisuuksista, tunnetilasta ja jopa sosiaalisesta taustasta. Käytännössä luonnollisen kaltaisen prosodian generoiminen laajan sanavaraston puhesynteesissä on vielä melko kaukainen tavoite, sillä prosodian mallintaminen on hyvin monimutkaista puuhaa. Erilaisia hierarkisia sääntöjä ajoituksen ja perustaajuuden säätelemiseen on kuitenkin toteutettu ja niillä on saatu aikaan jonkinasteisia parannuksia syntetisoidun puheen sujuvuuteen.

Puhesynteesi-lohkossa generoidaan lopulta itse puhesignaali. Tämä voidaan tehdä joko täysin parametriselta pohjalta, jolloin tuotetaan koneellisesti foneemien realisaatioita, tai sitten foneemien, difonien, trifonien tai muiden yksiköiden realisaatioita voidaan valita tietokannasta hakuprosessin kautta. Kummassakin tapauksessa nämä lyhyet puhepätkät liitetään yhteen ja näin syntyy lopullinen puhesignaali. Yhtenä suurimmista ongelmista synteesivaiheessa onkin jatkuvuuden varmistaminen pätkien liitoskohdissa, jotta vältytään häiriööäniltä ja vääristymiltä.

Seuraavaksi käydään läpi puhesynteesin kolme yleisintä toteutustapaa: *formanttisynteesi*, *konkatenaatio* ja *artikulatorinen synteesi*. Lisätietoja aiheesta löytyy mm. Sami Lemmetyn diplomityöstä (ks. kirjallisuusviitteet).

8.2.1 Formanttisynteesi

Tämä puhesynteesimenetelmistä vanhin oli pitkään hallitsevana tekniikkana syntetisaattorien toteutuksessa, mutta nykyään myös konkatenaatiosynteesi on hyvin yleinen toteutustapa. Formanttisynteesi perustuu jo tutuksi tulleeseen lähde-suodinmalliin, eli siinä generoidaan jaksollista ja ei-jaksollista herätesignaalia ja työnnetään se formantteja mallintavan resonaattorikytkennän eli ääntöväyläsuodattimen läpi. Menetelmä on siis periaatteeltaan hyvin yksinkertainen, mikä tekee siitä myös erittäin muokkaamiskelpoisen ja suhteellisen helpon toteuttaa. Sillä voidaan tuottaa mitä hyvänsä äänteitä, toisin kuin alla olevilla menetelmillä. Toisaalta herätesignaalin ja ääntöväylän epätäydellinen ja melko rajusti yksinkertaistava mallintaminen johtavat jokseenkin epäluonnolliselta kuulostavaan lopputulokseen.

Herätesignaaliksi riittää yksinkertaisimmillaan impulssijono tai sahalaita-aalto sekä kohinakomponentti, mutta puheen laadun ja ominaisuuksien säätömahdollisuuksien parantamiseksi on syytä käyttää mahdollisimman tarkkaa herätesignaalin mallia. Säädettäviä parametreja ovat yleensä ainakin perustaajuus, soinnillisen ja soinnittoman herätteen voimakkuudet sekä soinnillisuuden aste. Ääntöväylämallissa kuvataan tavallisesti jokaista formanttia napaparilla, jotta sekä formantin taajuus että sen kaistanleveys saadaan määriteltyä. Jotta puheesta tulisi ymmärrettävää, vaaditaan vähintään kolmen alimman formantin määrittämistä, mutta puheen laatu paranee jos formantteja otetaan vielä pari lisää. Kuten herätemallinkin parametreja, myös ääntöväylän taajuusvastetta kontrolloivia parametreja päivitetään jokaisen foneemin kohdalla. Ääntöväylämalli voidaan toteuttaa joko resonaattorien kaskadi- tai rinnakkaiskytkennällä. Molemmilla on omat etunsa ja haittansa erityyppisten äänteiden generoimisessa, mutta niitä ei tässä käydä läpi. Formantteja mallintavien suotimien lisäksi syntetisaattori voi sisältää herätesignaalia ja huulten energiasäteilyä mallintavat suotimet, sekä erillisen *antiresonaattorin* nasaaliäänteitä varten.

8.2.2 Konkatenaatio

Konkatenaatiosynteesi tarkoittaa ns. *leikkaa-liimaa*-synteesiä, jossa lyhyitä puhesegmenttejä valitaan ennalta äänitetystä tietokannasta ja liitetään peräkkäin haluttujen ilmaisujen aikaansaamiseksi. Periaatteessa todellisten puhesignaalien käyttäminen synteesin taustalla luo mahdollisuuden hyvinkin korkeaan laatuun, mutta käytännön rajoituksina tulee ensinnäkin vastaan tarvittavan muistikapasiteetin määrä. Mitä pidempiä puhesegmenttejä käytetään, sitä vähemmän syntetisoituun puheeseen tulee ongelmallisia segmenttien liitoskohtia, mutta samalla muistin tarve kasvaa. Toinen rajoittava tekijä on se, että ulostulopuhe on aina vahvasti riippuvainen tietokannassa olevasta puheesta, eli esim. puhujan henkilöllisyyttä tai tunnetilaa heijastavien ominaisuuksien muuttaminen on erittäin vaikeaa. Tiettyihin sovelluksiin menetelmä on yksitoikkoisuudestaan huolimatta täysin riittävä.

Minkä pituisia segmenttejä konkatenaatiossa sitten käytetään? Yleisimpiä ovat *foneemit* ja *difonit*, koska niillä saavutetaan riittävä joustavuus ilmaisujen muodostamisessa ja samalla pidetään muistikapasiteetin tarve kohtuullisena. Pidempien yksiköiden, kuten tavujen tai sanojen, käyttäminen on monestakin syystä mahdotonta tai järjetöntä. Difonien käyttö tarjoaa kohtalaisen hyvät mahdollisuudet koartikulaation huomioimiseen, nimittäin difoni sisältää kahden peräkkäisen foneemin välisen siirtymäajan sekä ensimmäisestä foneemista loppupuoliskon ja jälkimmäisestä foneemista alkupuoliskon. Difonikonkatenaatiossa segmenttien liitoskohdat siis osuvat kunkin foneemin keskikohtaan, jossa monet äänteet ovat tasaisimmillaan ja täten liitoskohtien vääristymien voidaan odottaa minimoituvan. Siinä missä erilaisia foneemeja tarvitaan synteesiä varten 40 - 50 kappaletta, tarvitaan difoneja 1500 - 2000 kpl, mutta tämän kokoinen tietokanta on usein vielä toteutettavissa (Lemmetty, 1999). Toisaalta foneemien käyttö on joustavin keino generoida erilaisia ilmauksia, ellei oteta huomioon sitä että tiettyjä äänteitä on suorastaan mahdotonta erottaa puheesta omiksi segmenteikseen (esim. klusiilit).

Sekä foneemien että difonien käytössä suurin haaste on peräkkäin liimattavien segmenttien yhteensopivuuden varmistaminen, mikä edellyttää vähintään perustaajuuden ja intensiteetin säätömahdollisuuksia. Luonnollisen prosodian luominen syntetisoituun puheeseen on nykypäivän menetelmillä vielä mahdotonta, joskin joitan yritelmiä on toki julkaistu. Yhtenä merkittävänä ongelmana konkatenaatiossa on vielä se, että tietokannan luominen on hyvin työlästä. Kaikki tarvittavat foneemit allofoneineen on saatava äänitettyä, ja tämän jälkeen ne täytyy vielä saada segmentoitua ja merkittyä asianmukaisilla ”nimilapuilla” (engl. *labeling*), jotta sopivan segmentin valitseminen tietokannasta olisi mahdollista. Toki tiettyjä operaatioita voidaan myös jossain määrin automatisoida.

8.2.3 Artikulatorinen synteesi

Tässä materiaalissa esitellyistä menetelmistä mallirakenteeltaan ja laskennallisilta vaatimuksiltaan ylivoimaisesti raskain puhesynteesitapa on artikulatorinen synteesi, jossa pyritään mallintamaan ihmisen puheentuottoa mahdollisimman täydellisesti. Hankalan toteutettavuutensa takia artikulatorinen synteesi ei ole vielä yleistynyt, eikä sillä ole saavutettu samanlaista menestystä kuin muilla synteesimenetelmillä, mutta periaatteessa sen mahdollisuudet luonnolliselta kuulostavaan syntetisoituun puheeseen ovat parhaimmasta päästä. Esimerkiksi äänteitä ja äänneyhdistelmiä, joita ihminen ei fysiologiansa perusteella yksinkertaisesti kykene tuottamaan, ei artikulatorisessa synteesissäkään tuoteta. Muissa menetelmissä tällaisten häiriööääninä havaittavien äänten tuottaminen sen sijaan on mahdollista. Lisäksi transienttiluonteisten äänitapahtumien luominen voidaan artikulatorisessa synteesissä tehdä muita menetelmiä tarkemmin.

Artikulatorisessa synteesissä mallinnetaan siis fysikaalisilla malleilla sekä ihmisen ääntöväylän että äänihuulten toimintaa. Ääntöväylän mallina toimii yleensä joukko pinta-alafunktioita, jotka kuvaavat kurkunpään ja huulten välillä esiintyviä ääntöväylän poikkipinta-alan muutoksia, eli periaate on samantapainen kuin akustisessa putkimallissa. Suurella määrällä erilaisia säätöparametreja säädellään yksityiskohtaisesti esimerkiksi huulten ja kielen asentoa, keuhkojen ilmanpainetta ja äänihuulten jännitystä. Mallintamisen perustana käytetty data hankitaan yleensä luonnollisen puhetapahtuman röntgenanalyysilla (Lemmetty, 1999), mutta tarkka mallintaminen on arvatenkin äärimmäisen monimutkaista.

8.3 Lisätietoa

Kaupallisista ja ei-kaupallisista puhesyntetisaattoreista ja niiden kehityshistoriasta on julkaistu useita erinomaisia yhteenvetoja, joiden yksityiskohtia ei tässä lähdetä toistamaan. Luetellusta (ja luettelon ulkopuolisesta) kirjallisuudesta sekä tietysti internetistä löytänee kukin haluamansa lisätiedot. Karkeana yhteenvetona sanottakoon kuitenkin, että puhesynteesin kehitys on monien tahojen mielestä nytkähtänyt liikkeelle 1930-luvulla, jolloin Bell Laboratories:ia edustanut Homer Dudley kehitti *Voder*-nimisen (Voice Operating Demonstrator) puhesyntetisaattorin, ja esitteli sen 1939. Se oli käsikäyttöinen laite, jossa valittiin kohinaluonteinen tai jaksollinen herätesignaali ja syötettiin se nipulle kaistanpäästösuotimia, joiden vasteita säädeltiin sormin. Puheen perustaajuus oli valittavissa pedaalilla. (Tätä ei kuka tahansa osannutkaan soittaa.) Voderin katsotaan olleen ensimmäinen osoitus siitä, että puhetta voidaan tuottaa keinotekoisesti, ja se innostikin monia tutkimusyksiköitä panostamaan puhesynteesiin. Varhaisimmat versiot formantti- ja artikulatorisesta syntetisaattorista ilmestyivät 50-luvulla. Digitaalitekniikan keksimisen jälkeen saatiin Japanissa aikaan ensimmäinen kokonaisena TTS-järjestelmänä pidetty englanninkielinen systeemi vuonna 1968. Se perustui artikulatoriseen malliin ja sen tuottaman puheen väitetään olleen melko ymmärrettävää, joskin monotonista (Huang, Acero, Hon, 2001).

Formanttisynteesi, josta esimerkkinä kuuluisa *Klattalk*, hallitsi puhesynteesin kenttää melko pitkään, mutta *PSOLA*-menetelmän (Pitch-Synchronous Overlap-Add) kehittäminen vuonna 1985 toi huomattavaa edistystä konkatenaatiosynteesin kehitystyöhön. PSOLAssa ideana on erottaa puhesignaalista kehyksiä pitch-jakson välein ja synteesivaiheessa summata tällaisia osittain päällekäin osuvia kehyksiä yhteen siten, että haluttu ulostulosignaalin aikaskaala ja pitch toteutuvat. Näin voidaan muuttaa erilaisia puheen prosodisia ominaisuuksia toisistaan riippumatta, eli esimerkiksi puhenopeutta voidaan pienentää perustaajuuden pysyessä ennallaan. Tällöin tiettyjä analyysikehyksiä kopioidaan sen verran, että haluttu, alkuperäistä pidempi aika-akseli täyttyy, mutta kehykset sijoitetaan synteesissä edelleen alkuperäisen pitch-jakson välein, jolloin perustaajuus ei muutu. Vastaavasti pitchiä voidaan korottaa asettelemalla synteesikehyksiä tiheämpään kuin alkuperäisessä signaalissa. Kehysten eliminoiminen vastakkaisten efektien luomiseksi on tietysti yhtä lailla mahdollista. Arvatenkin PSOLA sopii parhaiten soinnilliselle puheelle, josta pitch-jaksot ovat määritettävissä. PSOLA on itse asiassa erittäin herkkä pitch-estimaatin virheille, mikä käytännössä aiheuttaa usein ongelmia. PSOLAn vahvuus taas on sen äärimmäisessä yksinkertaisuudessa, joten laskennallisesti se on kevyt toteuttaa. Aikatason PSOLAssa kehyksiä ei edes muuteta analyysin ja synteesin välillä, vaan kaikki efektit luodaan kehysten lukumäärää ja välimatkoja säätelemällä. Valmiita PSOLAn Matlab-toteutuksia voi imuroida vaikkapa Digital Audio Effects -kirjan kautta sivulta

`http://www.unibw-hamburg.de/EWEB/ANT/dafx2002/DAFX_Book_Page/matlab.html`

jossa Kappale 7 sisältää PSOLAn. Erilaisia variaatioita perus-PSOLAsta käytetään yleisesti konkatenaatiosynteesissä.

Myös mallia, jossa puhesignaali esitetään harmonisten sinikomponenttien ja kohinakomponentin summana, on jonkin verran yritetty käyttää puhesynteesissä. Oikeastaan tämä menetelmä sopii silti paremmin laulusynteesiin kuin puheeseen. Muita puhesynteesiin sovellettuja signaalinkäsittelymenetelmiä ovat mm. HMM:t (Hidden Markov Model) ja neuroverkot. Kumpiakin on menestyksekkäästi sovellettu puheentunnistukseen, mutta puhesynteesissä niillä uskotaan olevan vielä löytämättömiä voimavaroja.

Loppujen lopuksi on vielä korostettava, että puhesynteesi koskettaa lukuisia eri tutkimuksen osa-alueita. Vaikka pelkästään signaalinkäsittelymenetelmien tutkiminenkin työllistää sankan joukon puhesynteesin tutkijoita, tarvitaan entistäkin parempiin puhesyntetisaattoritoteutuksiin pyrittäessä yhteistyötä ainakin fonetiikan, kieliopin, sanasto-opin, semantiikan, pragmatiikan, tiedonhaun, ohjelmistotekniikan ja signaalinkäsittelyn aloilla.

Kirjallisuutta

- X. Huang, A. Acero, H.-W. Hon, Spoken Language Processing, Prentice Hall PTR, 2001.
- S. Lemmetty, Review of Speech Synthesis Technology, Master’s Thesis, Helsinki University of Technology, 1999.
- U. Zölzer (editor), DAFX - Digital Audio Effects, John Wiley & Sons, Ltd, 2002.