Puhesynteesi

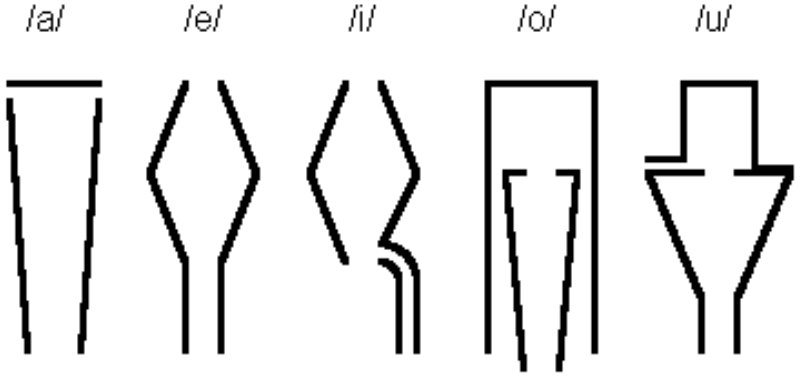
Martti Vainio

Fonetiikan laitos, Helsingin yliopisto

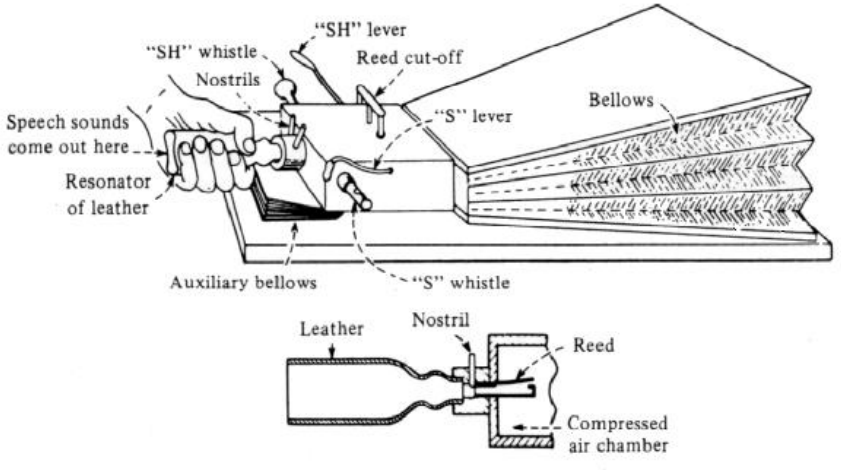
Puhesynteesin historiaa

- Mekaaniset synteesit: 1700-luvulla asiaa harrastivat Wolfgang von Kempelen ja Christian Kratzenstein.
- 1900-luvulla tulivat elektromekaaniset sekä elektroniset synteesit ja vuosisadan loppupuolella digitaaliset syntisaattorit.
- Ks. `http://www.acoustics.hut.fi/ slemmett/dippa/chap2.html`

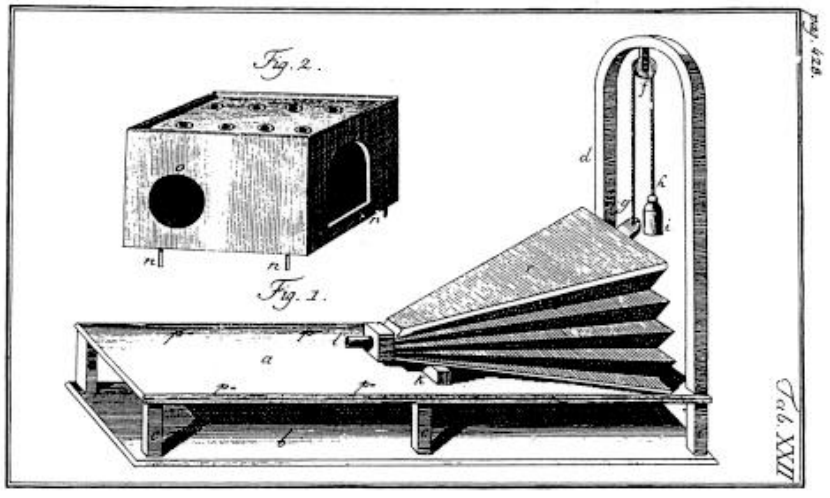
Historiaa: Kratzenstein



Historiaa: Kempelen



Historiaa: Kempelen



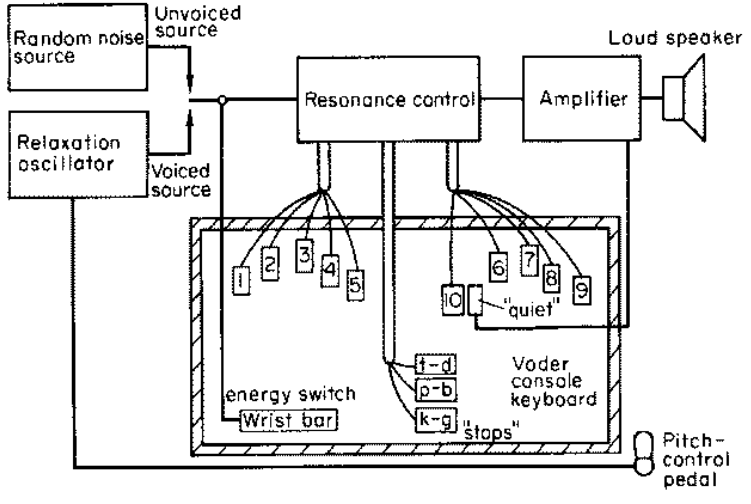
Puhesynteesin historiaa: 1930-luku

- 1936: Englannin puhelinyhtiön puhuva kello käytti optista tallennusta – lausekkeet, sanat ja sanojen osat.
- 1939: Bell Laboratorion VODER (Homer Dudley) – mekaaninen urkujen kaltainen laite jolla voitiin ’soittaa’ puhetta. (1)*
- Dudleyn VOCODER, jossa puhesignaali jaettiin lähde-suodin mallin mukaisesti.

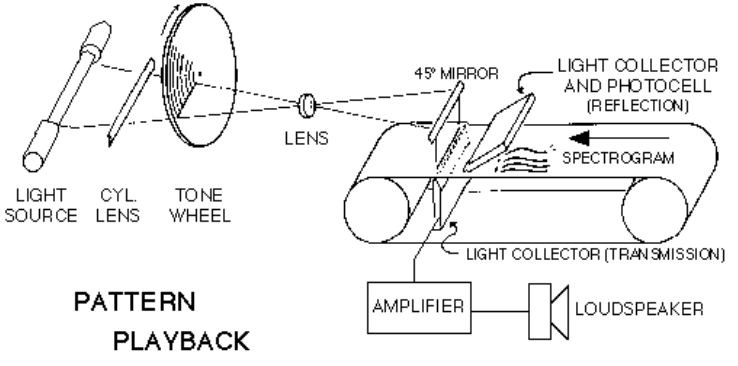
* = ääniesimerkin numero osoitteessa

`http://www.festvox.org/history/klatt.html`

Historiaa: VODER



Historiaa: pattern playback



Historiaa: 1940 ja 1950-luvut

- Terminaalianalogiaan perustuvat mallit – formanttisynteesi
- Gunnar Fantin OVE, 1953. (4)
- Georg Rosenin artikulatorinen DAVO-syntetisaattori (MIT, 1958). (11)

Historiaa: 1960-luku

- Ensimmäiset digitaaliset mallit ja sääntösynteesi – tekstistä puheeksi
  - Brittienglanti; Holmes, Mattingly ja Shearme, 1964. (17)
  - Cokerin sääntöpohjainen artikulatorinen malli, 1968. (19)
  - Mattinglyn ensimmäinen prosodinen synteesi, 1968. (20)

Historiaa: 1970-luku

- TTS (Text-To-Speech) tuotteet ja difonisynteesi
  - Ensimmäinen täydellinen TTS-järjestelmä, Noriko Umeda, Japani, 1968. (24)
  - Lausetason fonologiset säännöt, Dennis Klatt, 1976. (21)
  - Lineaariprediktioon perustuvien difonien konkatenaatio, Joseph Olive, 1977. (22)
  - Votraxin Type-n-Talk, Richard Cagnon, 1978. (28)
  - MIT:n MITalk, Jonathan Allen, Sheri Hunnicut ja Dennis Klatt, 1979. (30)

Historiaa: 1980-luku

- Konkatenaatio valtaa alaa – suuremmat järjestelmät:
  - AT&T Bell Laboratories, TTS-järjestelmä, 1985. (34)
  - DECtalk (35)
  - DECtalk, 300 sanaa/minuutti. (36)

Historiaa: 1990-luku

- Tuotteet, monikielisyys, ’unit selection’
  - Yleinen ’unit selection’, CHATR, Japani, 1994.
  - Monikielinen MBROLA, ’vapaa’ synteesi, Belgia, 1995.
  - Mikropuhe, TIMEHOUSE, Suomi
- 2000-luku: ...
- Toisaalta kaupalliset järjestelmät perustuvat usein valmiiksi äänitettyyn materiaaliin ja sanojen liimaamiseen (vertaa 1936!) koska lopputulos on parempi.

Puhesynteesin kolme peruslajia:

1. Analyysi-resynteesi
   - LPC-synteesi
   - GSM koodaus . . .
2. Tekstistä puheeksi (TTS = Text-to-Speech)
   - Vammaissovellukset
   - Puhelinpalvelut; sähköpostin luku . . .
3. Konseptista puheeksi (CSS = Concept-to-Speech Synthesis)
   - Tietokantojen luku, listat, aikataulut
   - Dialogijärjestelmät

Kolme perusparametriä:

1. Sanaston suuruus
   - Rajattu sanasto – kuulutukset
   - Rajaton sanasto – vapaa teksti
2. Synteesitapa
   - Valmiin puheen *leikkaa-liimaa* menetelmät
   - Pienten yksiköiden konkatenaatio
   - Formanttisynteesi
3. Syötteen laatu
   - Puhe
   - Teksti
   - Tietokanta

Kahdenlaista motivaatiota:

1. Sovellukset
   - Vammaissovellukset
   - TTS-järjestelmät
   - Dialogijärjestelmät
2. Tieteellinen tutkimus
   - Puheen havaitseminen – kontrolloidut ärsykkeet
   - Puheen tuoton mallit
   - Prosodian tutkimus

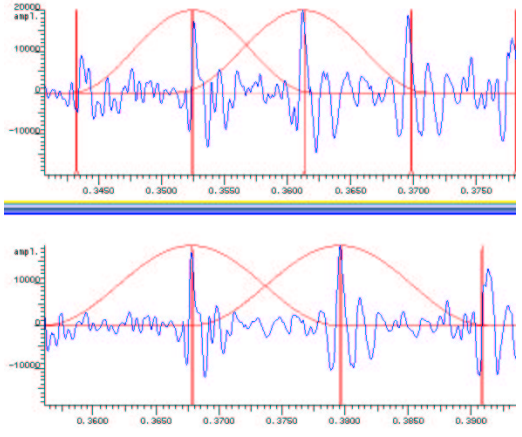
Kolme lähestymistapaa:

1. Konkatenaatio
   - Sanat, lausekkeet, lauseet
   - Sanaa pienemmät osat; tavut, puolitavut
   - Difonit
   - ”Mikrofoneemit"
2. Formanttisynteesi
   - Puhe tuotetaan synteettisesti alusta pitäen
3. Artikulatorinen synteesi
   - Fyysiset mallit – puhe on fysiikkaa

Konkatenaatio

- Oikeata puhetta leikkaa-ja-liimaa -periaatteella.
- Mitä leikataan: lausekkeita, sanoja, tavuja, puoli-tavuja, äänteitä, difoneja.
- **Miten:** tarkasti leikatut yksiköt voidaan liimata päistään yhteen, tasoitus (smoothing), PSOLA (pitch-syncronous overlap and add) . . .
- **Etuja:** äärellinen määrä puhedataa riittää, prosessointi on yksinkertaista, lopputuloksena korkeatasoinen ääni.

Difonikonkatenaatio: TD-PSOLA

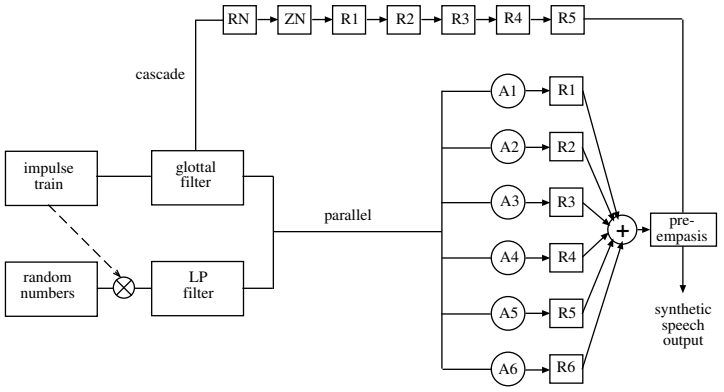


- TD-PSOLA: alennettu perustaajuus

Formanttisynteesi

- **Miten:** Generoidaan periodista ja aperiodista ääntä ja niitä yhdistelemällä tuotetaan puheenkaltainen ääni.
- **Etuja:** erittäin muokkautuvainen, voidää päästä lähes täydelliseen lopputulokseen, suhteellisen helppo implementoida, tieteellisesti kiinnostava.

Formanttisynteesi: kaavio



- Klatt syntetisaattori

Formanttisynteesi: rinnakkainen vai sarjassa

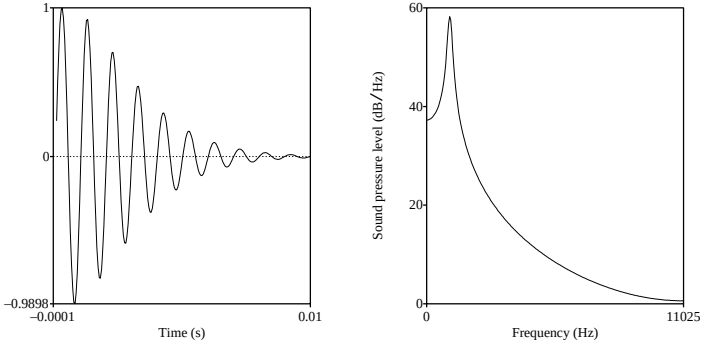
- Sarjaan kytkettyjen resonaattorien etu on, että niille täytyy kertoa vain formanttien taajuudet ja kaistanleveydet; amplitudit hakeutuvat automaattisesti oikeiksi.
- Rinnakkaisesta synteesiä rasittaa lisäksi vaatimus formanttien amplitudeista; rinnakkaismallilla voidaan kuitenkin simuloida helpommin sellaisia konsonantteja (etenkin frikatiiveja), joilla on antiformantteja. (Toisaalta rinnakkaismallin resonanssien summaaminen implikoi sitä, että resonaattorit ovat itsenäisempiä ja ovat siten itsenäisesti kontrolloitavissa.)
- Sarjamalli perustuu suoremmin puheentuoton akustiseen teoriaan, jonka mukaan ääntöväylän siirtofunktio on esitettävissä suotimien tulona.

Formanttisynteesi: glottaalinen eksitaatio

- Periaatteessa pelkkä impulssijono riittää tuottamaan puheen kaltaisen tuloksen syntetisaattorista. Luonnollisuus vaatii kuitenkin lähteeltä enemmän. Esim. KLATT-synteesissä lähteeseen liittyy useita parametrejä, joiden avulla voidaan mallintaa muutoksia niin eri äänteiden kuin puhujienkin välillä.
- Glottislähteen parametrejä ovat mm. sulkeuma- ja avaumavaiheiden suhde (open quotient), aspiraatiohälyn määrä, ns. jitter (perustaajuuden perturbaatio) ja lähteen spektraalinen kaltevuus. Myös glottiksen alapuolisen väylän vaikutus pulssin muotoon on otettu huomioon.

Formanttisynteesi: resonaattori

- Formanttiresonaattorin impulssivaste ja sen spektri; formantin kaistanleveys on suoraan verrannollinen impulssivasteen vaimenemiseen.

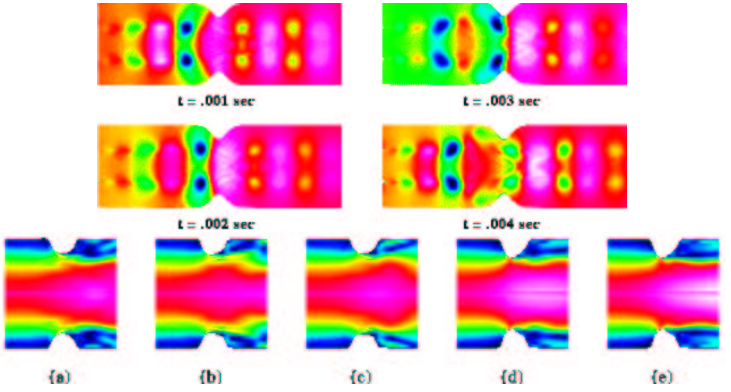


Artikulatorinen synteesi

- **Miten:** Mallinnetaan ääniväylää pinta-alojen ja ilman virtausten sekä heijastusten suhteen – puheentuotto nähdään sovellettuna fysiikkana.
- **Etuja:** Parantunut kontrolli, potentiaalisesti luonnollista puhetta, perustutkimusta.

Artikulatorinen synteesi: esimerkki

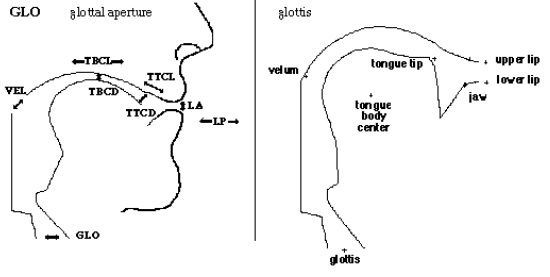
- Äänenpaineet ja hiukkasnopeudet artikulatorisessa mallissa:



Artikulatorinen synteesi: esimerkki 2

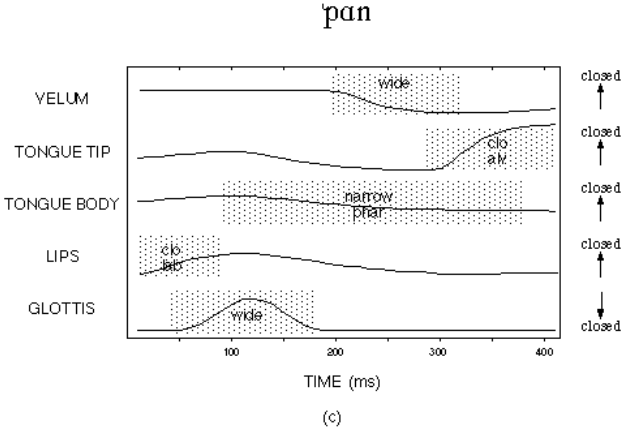
- Haskins laboratorion artiulatorinen malli:

	tract variable	articulators involved
LP	lip protrusion	upper & lower lips, jaw
LA	lip aperture	upper & lower lips, jaw
TTCL	tongue tip constrict location	tongue tip, tongue body, jaw
TTCD	tongue tip constrict degree	tongue tip, tongue body, jaw
TBCL	tongue body constrict location	tongue body, jaw
TBCD	tongue body constrict degree	tongue body, jaw
VEL	velic aperture	velum
GLO	glottal aperture	glottis

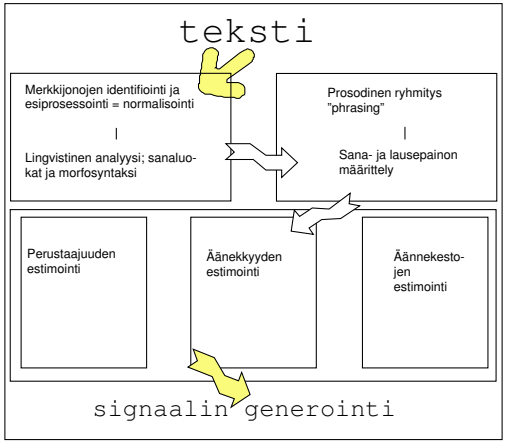


Artikulatorinen synteesi: esimerkki 3

- Haskins laboratorion artiulatorinen malli:



Tekstistä puheeksi:



Modulaarisuus

- Lähes kaikki modernit puhesynteesijärjestelmät ovat modulaarisia: tekstianalyysiä seuraa prosodiset modulit, joita seuraa synteesimodulit.
- Usein suuremmat modulit on vielä jaettu useimpiin tarpeen mukaan – esim. intonaatiota voidaan mallintaa usealla tavalla saman järjestelmän sisällä.

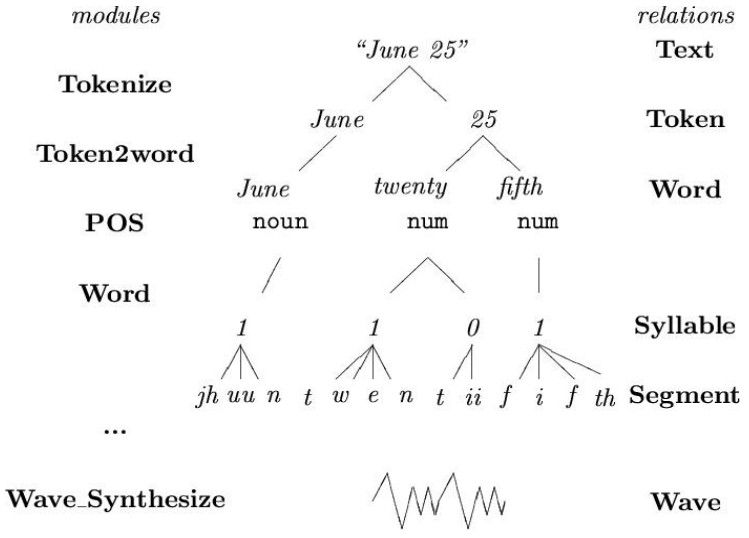
Data vs. tieto

- Historiallisesti sääntösynteesijärjestelmät ovat perustuneet tietoon – datapohjaiset järjestelmät ovat uudempi suuntaus.
- Kielen kombinatorinen kompleksisuus on kuitenkin niin valtava, että suuretkin tietokannat ovat tuomittuja edustamaan vain äärimmäisen pientä osaa koko puhutun kielen avaruudesta.

Tekstin analyysi

- Tekstin analyysiin kuuluu kaikki tekstin esiprosessointi ja normalisointi.
- Teksti muunnetaan järjestelmän ymmärtämään lingvistiseen muotoon, joka sisältää yleensä sanat ja niiden kieliopilliset kategoriat, morfologiset analyysit, fonologiset transkriptiot, aksentuaaliset ja tonaaliset piirteet sekä prosodisten rajojen paikat.

Tekstin analyysi: esimerkki Festivalista



Prosodinen esiprosessointi

- Prosodinen esiprosessointi pitää sisällään syntaktisen analyysin (joka voi yksinkertaisimmillaan olla funktiosanojen tunnistamista) ja lauseiden sekä lausekkeiden rajojen paikantamisen.
- Myös lausepainon paikan määritys kuuluu prosodiseen tähän vaiheeseen.

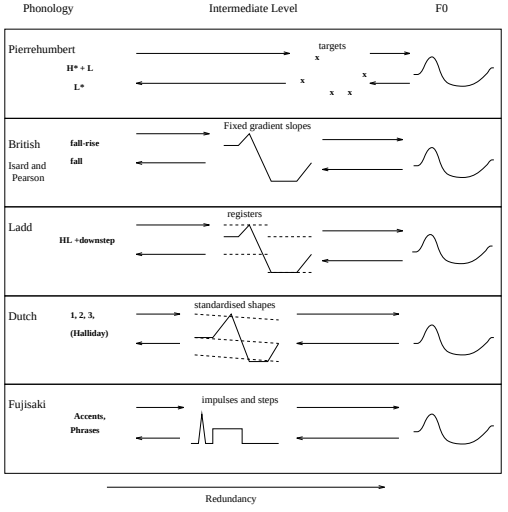
Prosodian estimointi: ajoitus

- Järjestelmään sisältyvän ajoituskomponentin tehtävänä on laskea puheelle sen temporaalinen rakenne annetusta symbolisesta syötteestä: foneemit, paino- ja lausepainomerkinnät.
- Yleensä ajoituksella tarkoitetaan äännekestoja, mutta muunlaistakin temporaalista informaatiota tarvitaan; esim. perustaajuuden huippujen paikka vokaaliin nähden.
- Ajoitus voidaan laskea joko sääntöjen avulla tai dataan perustuen esim. keinotekoisia hermoverkkoja käyttäen.

Prosodian estimointi: intonaatio

- Intonaatiokomponentin tehtävänä on laskea tuotettavalle lauseelle sen perustaajuuskontuuri ajoituskomponentin käyttämästä syötteestä ja sen tuottamista äännekestoista.
- Teorioiden ja mallien suhteen intonaation tutkimus on äärimmäisen vaihtelevaa ja mallien kirjo heijastuukin synteesijärjeselmiin.
- Fonologisella puolella ei ole kunnollista konsensusta yksiköiden suhteen ja foneettisella puolella ei ole yksimielisyyttä siitä, miten käyrät tulisi laskea: lauseke ja aksenttikomponenttien superpositio (Fujisaki), tonaaliarvojen interpolaatio (Pierrehumbert), linjasegmenttien konkatenaatio (IPO).

Intonaatiomallit:



Signaalin generointi

- Synteesikomponentti ottaa vastaan äännejonotietoa ja prosodista informaatiota, joista sen tehtävänä on generoida kuultava signaali.
- Historiallisesti signaalin generointi on perustunut lähde-suodin -malliin; formanttisynteesi. Nykyisin kuitenkin suurin osa syntetisaattoreista käyttää jonkinasteista konkatenaatiomenetelmää (difoni tai ns. unit-selection).
- Difonikonkatenaatiossa yksikköinä ovat nimen mukaisesti difonit (kahden äänteen keskipisteiden välinen osa). “unit selection” -tyyppisessä synteesissä yksikön koko vaihtelee jopa kokonaisista lauseista difoniin.