Puheteknologian perusteet
Luento 2 – Sovellukset

Kurssilla “Informaatioteknologian perusteet”

Tom Bäckström

28.2.2018

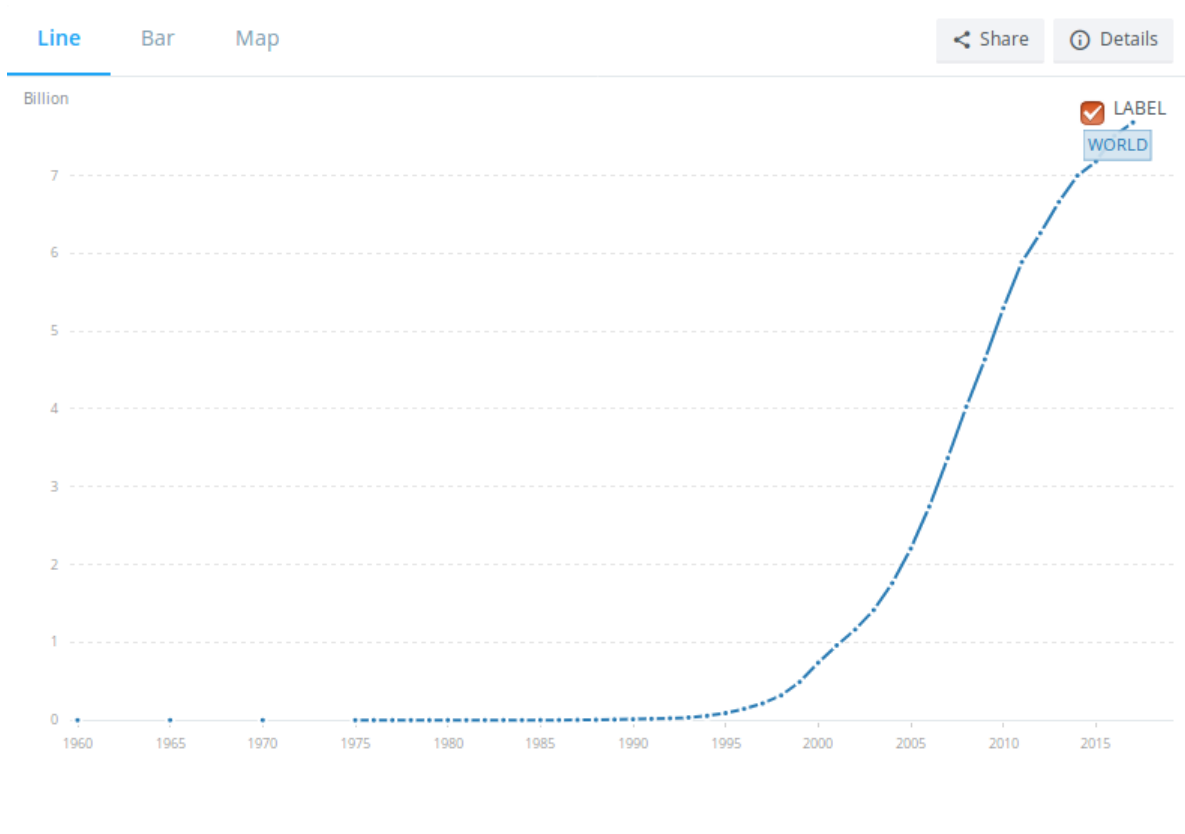
Puheteknologian sovelluksia

- Puheenkoodaus
- Puheensiistaus/-ehostus
- Puheentunnistus
- Puhesynteesi

Puheenkoodaus

Puheteknologian sovelluksia

Digitaalinen puheensiirto (=kännykät) oli lähtölaukaus mobiilin informaatiotekniikan vallankumoukselle 90-luvulla.



Puheenkoodaus

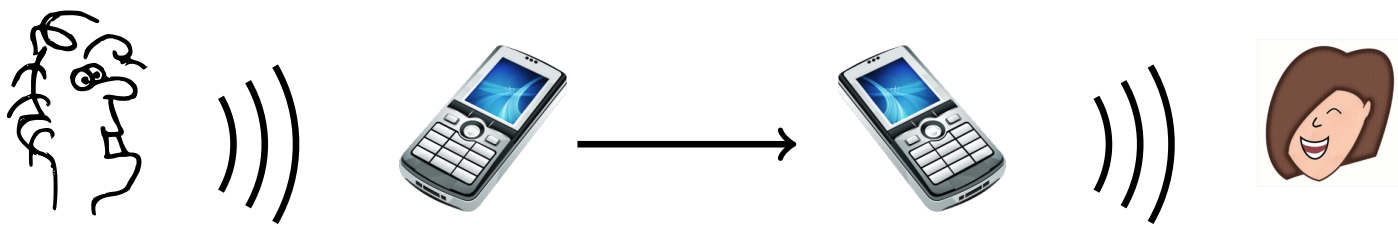
Puheteknologian sovelluksia

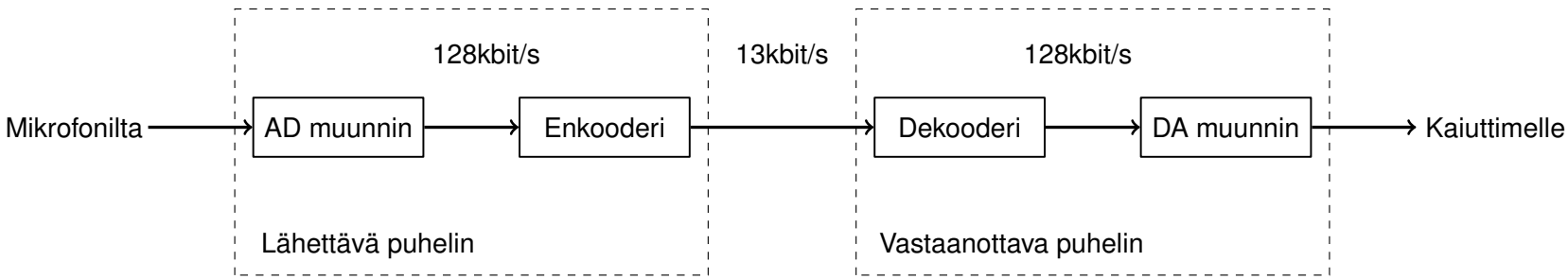
- Puheenkoodauksella tarkoitetaan puheäänen digitaalista pakkaamista mahdollisimman pienen kokoon (bit/s) siten että se voidaan silti rekonstruoida hyvällä äänenlaadulla.
  - Tärkein sovellus on telekommunikaatio (puhelimesta-puhelimeen tai puhelimesta pilveen)
  - Myös puheäänen tallennus
- Tavoitteet ovat ristiriitaiset
  - Ääni pitäisi pakata mahdollisimman pieneen tilaan, käyttäen mahdollisimman vähän CPU-kapasiteettia ja muita resursseja.
  - Äänenlaadun pitäisi olla vastaanottajalla mahdollisimman hyvä (ymmärrettävyys, miellyttävyys, ei kohinaa), sekä siten ettei pakkaaminen häiritse vuorovaikutusta (ei viivettä).
  - Laatua voi yleensä aina parantaa jos lisää resursseja tai viivettä.

Puheenkoodaus

Puheteknologian sovelluksia

Koodauksen perusidea





Puheenkoodaus

Puheteknologian sovelluksia

- Puheenkoodauksessa pakataan signaali tyypillisesti 10%:n kokoon alkuperäisestä.
- Puheenkoodaus on käytännössä aina häviöllistä pakkausta (lossy coding).
  - Tarkoituksena ei ole rekonstruoida samaa signaalia, vaan signaali joka *kuulostaa* samalta.
  - Yksityiskohdat, joita korva ei voi erottaa, voidaan jättää epätarkaksi.
  - Mielenkiintoista on että korva huomaa helposti lisäykset ja poistot, mutta huonommin äänen korvaamista toisella.
  - Kohinaisia yksityiskohtia voidaan korvata “millä tahansa” kohinalla, mutta niitä ei voi poistaa.

Puheenkoodaus

Puheteknologian sovelluksia

Esimerkkiääniä

Alkuperäinen 16 bit/sample (128 kbit/s)
Kvantisointi 8 bit/sample (64 kbit/s)
Kvantisointi 4 bit/sample (32 kbit/s)
Kvantisointi 2 bit/sample (16 kbit/s)
Kvantisointi 1 bit/sample (8 kbit/s)
GSM-pakattu (AMR-NB) (13 kbit/s)

Puheenkoodaus

Puheteknologian sovelluksia

- Perinteinen lähestymistapa puheenkoodaukseen tunnetaan nimellä *Code-Excited Linear Prediction (CELP)*.
  - Se perustuu puheentuottomalliin (kts. edellinen luento) sekä kuulon mallintamiseen.
- Puheentuottomallia kutsutaan lähde-suodin malliksi.
  - Puheen lähde (source tai excitation) on periodinen (soinnilliset äänteet) tai kohinainen (soinnittomat äänteet).
  - Äänivaylän vaikutusta mallinetaan suodattimella (linear prediction). Tätä suodatinta voi verrata ekvalisaattoriin, joka korostaa joitain taajuusalueita ja vaimentaa toisia.
  - Lähde-suodin malli on puheentuottomallina *hyvin* karkea aproksimaatio. Sitä ei pidä tulkita fysikaaliseksi malliksi, koska mallin parametrit eivät helposti käänny fysikaalisesti järkeenkäyviksi äänivaylän muodoiksi.
  - CELP mallia käytetään koska se on yksinkertainen ja se on verrattain helppo ohjelmoida, ja sillä voi tehokkaasti koodata ääntä joka kuulostaa luonnolliselta.

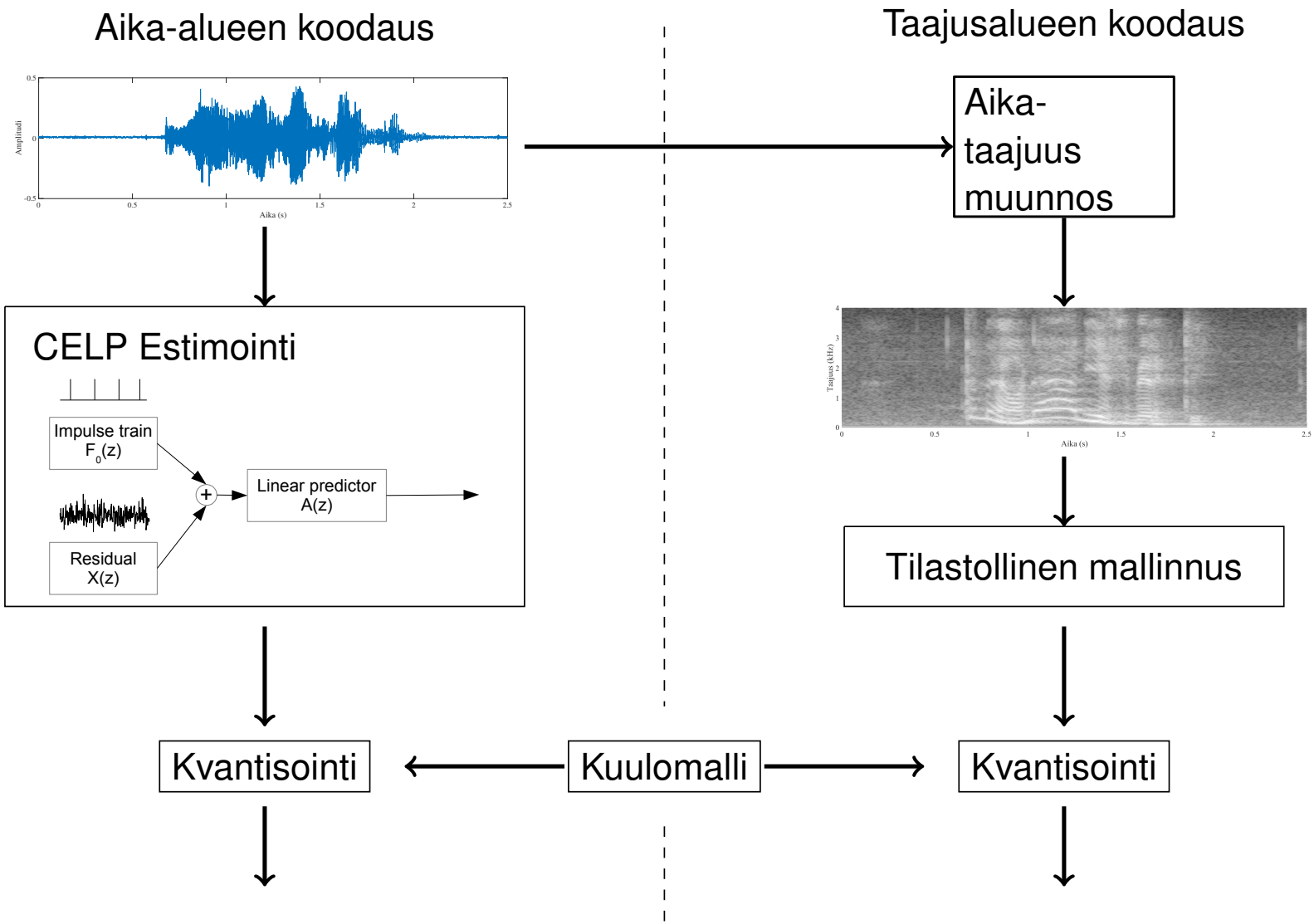
Puheenkoodaus

Puheteknologian sovelluksia

- CELP koodaus on aika-alueen koodausta.
  - Siinä mallinetaan aikasignaalia (kvantisoidaan aikasignaalin näytteitä).
- Audio-koodauksessa (tyyppiä MP3, AAC, Ogg, jne.) käytetään sen sijaan taajuusalueen koodausta.
  - Taajuusalueen koodauksessa kvantisoidaan “spektrogrammia” (kts. edellinen luento).
  - CELP olettaa että ääniä on vain yksi, kun musiikissa on tyypillisesti monta ääntä päällekkäin. Siksi musiikki ei toimi hyvin CELP koodekilla.
  - Taajuusalueen metodit ovat laskennallisesti yksinkertaisempia (What-You-See-Is-What-You-Get), kun CELP-metodeissa joudutaan käyttämään raskaita matriisi-operaatioita.
- Taajuusalueen koodaus on yleistymässä myös puheenkoodauksessa.
  - Enää ei tarvitse välittää onko ääni puhetta vai musiikkia.
  - CELP on tehokkaampi (ainakin vielä) $\Rightarrow$ Pieni tehokkuustappio.
  - Yksinkertaistaa ohjelmarakennetta ja kuulon mallintamista.

Puheenkoodaus

Puheteknologian sovelluksia



Puheenkoodaus

Puheteknologian sovelluksia

Puheenkoodauksen lyhyt historia

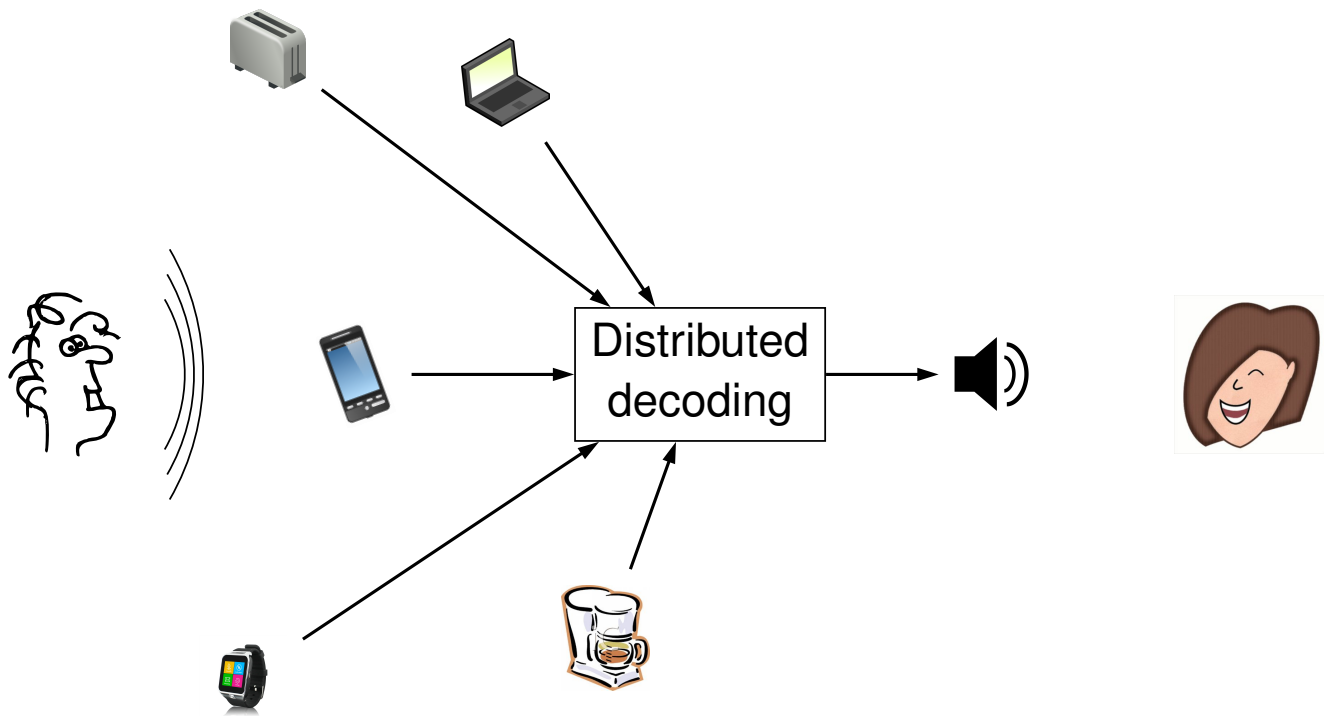
- **1G** Analogiset NMT puhelimet (1980-luvun lopulla)
- **2G** Digitaaliset GSM puhelimet (ensimmäisenä Radiolinja, 1991)
- **3G** Nopeampi verkko (1998)
- **AMR-WB** Parempaa äänenlaatua (valmis vuonna 2002, mutta käyttöönotto on ollut/oli hidasta).
- **4G** Vielä nopeampi verkko (2008)
- **EVS** Vielä parempaa äänenlaatua ja yhteensopivuus IP-verkkojen kanssa (valmis 2014, käyttöönotto alkanut)
- **5G** ?

Puheenkoodaus

Puheteknologian sovelluksia

Puheenkoodauksen tulevaisuus / työn alla juuri nyt

- Hajautettu puheen ja äänen koodaus (distributed speech and audio coding)
- Tavoitteena on että kaikki laitteet joissa on mikrofoni lähettäisivät ääntä yhteisvoimin.



Puheenkoodaus

Puheteknologian sovelluksia

Puheenkoodauksen tulevaisuus / työn alla juuri nyt

- Monen mikrofonin käyttäminen parantaa äänenlaatua.
- Käyttäjä voi unohtaa missä puhelin on
  - Lähin laite nappaa äänen.
  - Käyttöliittymä helpottuu.

Haasteita

- Yksityisyys ja turvallisuus! (ks. avoin työpaikka)
  - Kaikki laitteet kuuntelevat koko ajan!
  - Eettinen insinööri voi toimia vain jos yksityisyys turvataan.
- Tehokkuus
  - Enkooderin pitää olla yksinkertainen jotta kaikki laitteet voivat jatkuvasti koodata (energiankulutus).
  - Pitää varmistua siitä että kaikki laitteet lähettävät uniikkia informaatioita. Ylimääräisistä mikrofoneista ei ole hyötyä jos ne lähettävät täsmälleen samaa signaalia.

Puheensiistaus/-ehostus

Puheteknologian sovelluksia

- Puhesovelluksia käytetään oikeassa elämässä
  - = ihan aina ei istutat äänieristetyssä studiossa.
- Taustaäänet ja huonekaiku ovat merkittäviä häiriöitä.
  - Kuvittele puhuvasi puhelimeen diskossa (taustamelu) tai vessassa (kaiku).
- Puheensiistauksella ja -ehostuksella viitataan menetelmiin jolla parannetaan äänenlaatua.
  - Kohinanpoisto-menetelmät (noise attenuation) poistavat häiritseviä ääniä.
  - Tilaäänenehostus-menetelmät (spatial enhancement) poistavat huonekaikua.

Puheensiistaus/-ehostus

Puheteknologian sovelluksia

Kohinanpoiston perusidea

- Oletetaan että kohina v on puhesignaalille s additiivista, siten että havainto (=mikrofonisignaali) on

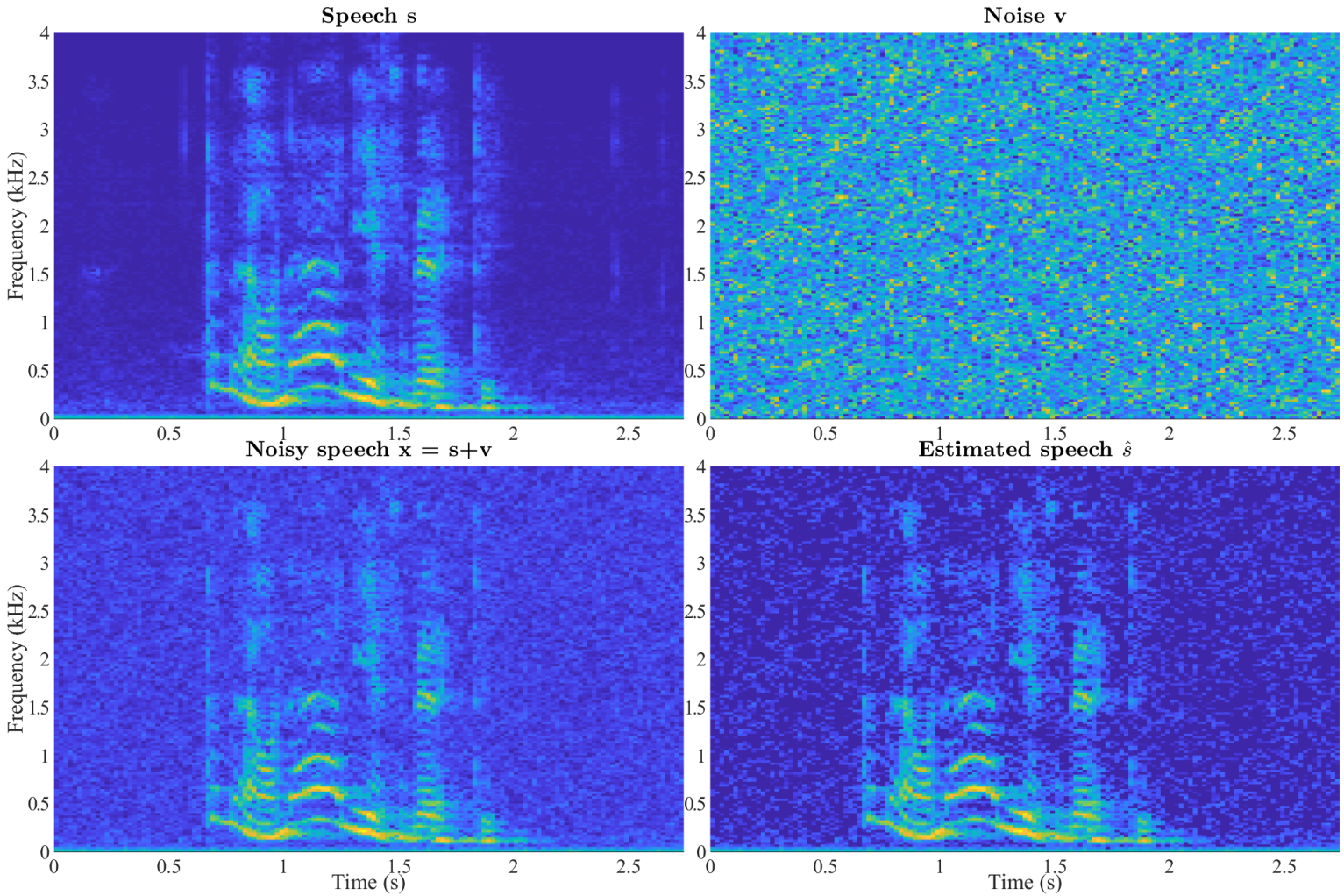
$$x = s + v.$$

- Kohinan energiaa $|v|^2$ on mahdollista estimoida, eli tehdään arvaus $|\hat{v}|^2$ siten että toivottavasti $|\hat{v}|^2 \approx |v|^2$.
- Koska $|x|^2 \approx |s|^2 + |v|^2$, on $|s| \approx \sqrt{|x|^2 - |\hat{v}|^2}$.
- Muuttujat x, s ja v ovat kompleksiarvoisia. Tehdään arvaus että v on pieni, joten kompleksitason kulmat ovat samoja $\frac{s}{|s|} = \angle s = \angle(x - v) \approx \angle x = \frac{x}{|x|}$.
- Seuraa että voidaan arvata puhesignaali

$$s = |s| \cdot \angle s \approx \frac{x}{|x|} \sqrt{|x|^2 - |\hat{v}|^2}.$$

Puheensiistaus/-ehostus

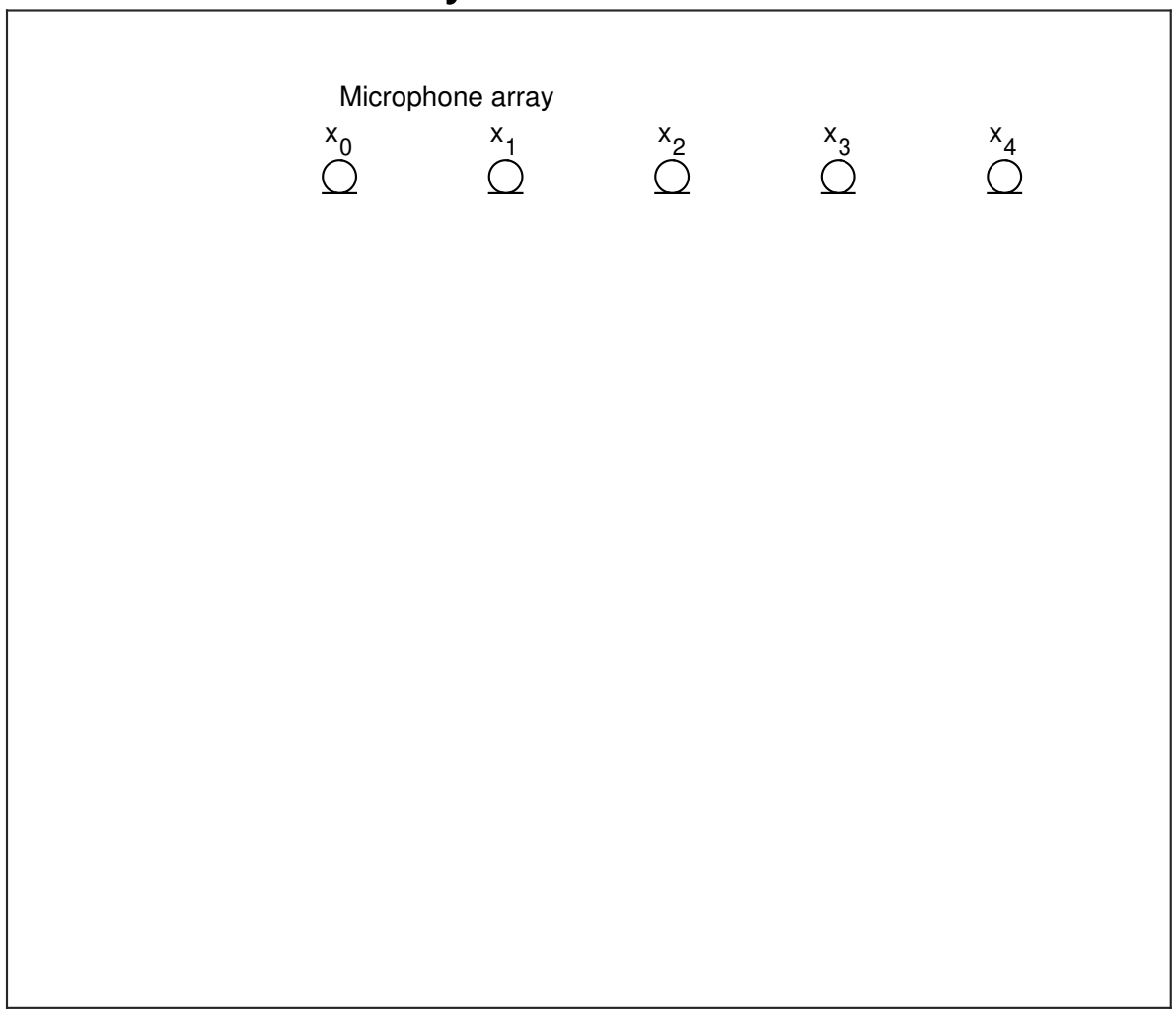
Puheteknologian sovelluksia



Puheensiistaus/-ehostus

Puheteknologian sovelluksia

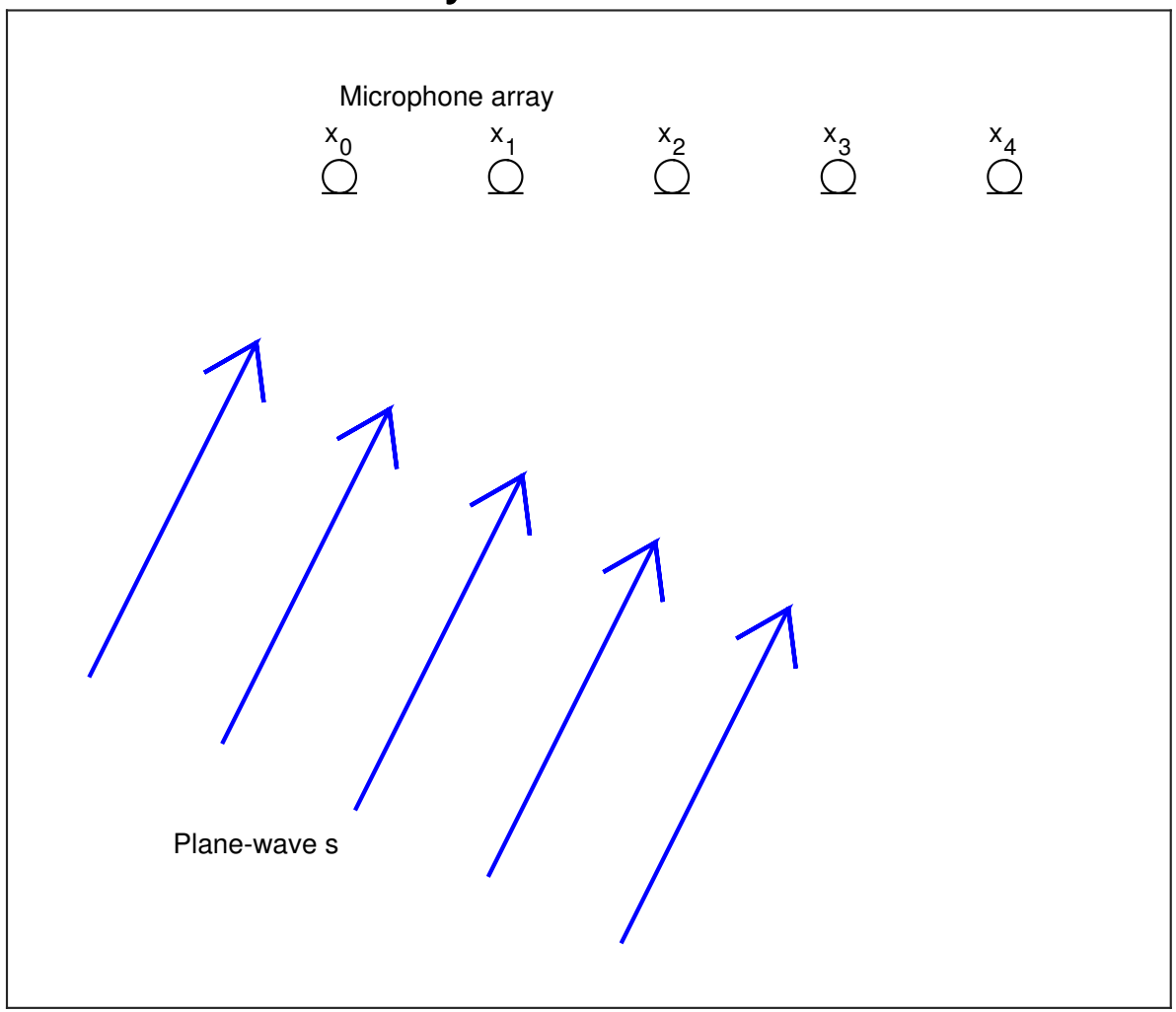
Mikrofonihilan käyttö tilaäänenehostamisessa



Puheensiistaus/-ehostus

Puheteknologian sovelluksia

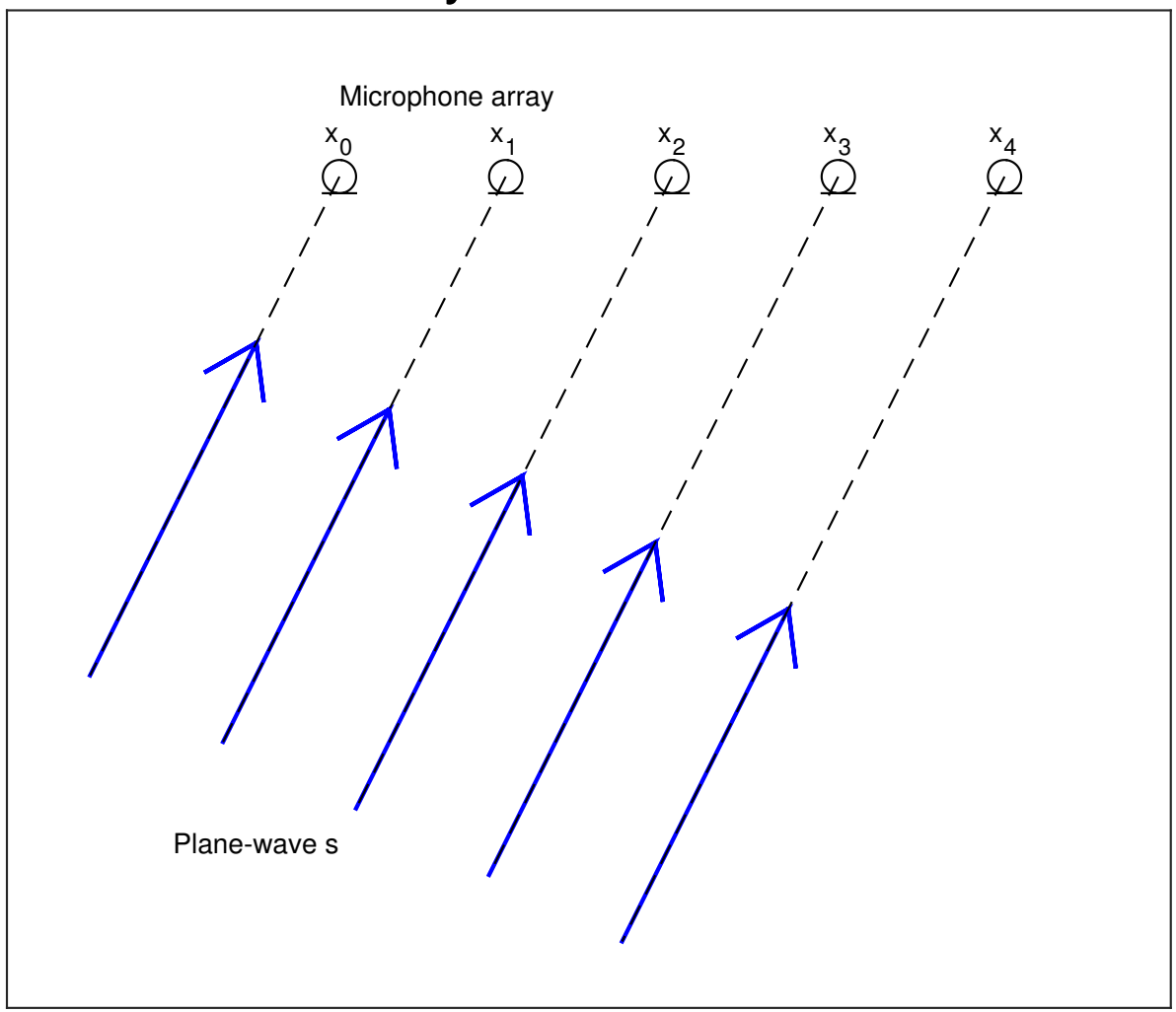
Mikrofonihilan käyttö tilaäänenehostamisessa



Puheensiistaus/-ehostus

Puheteknologian sovelluksia

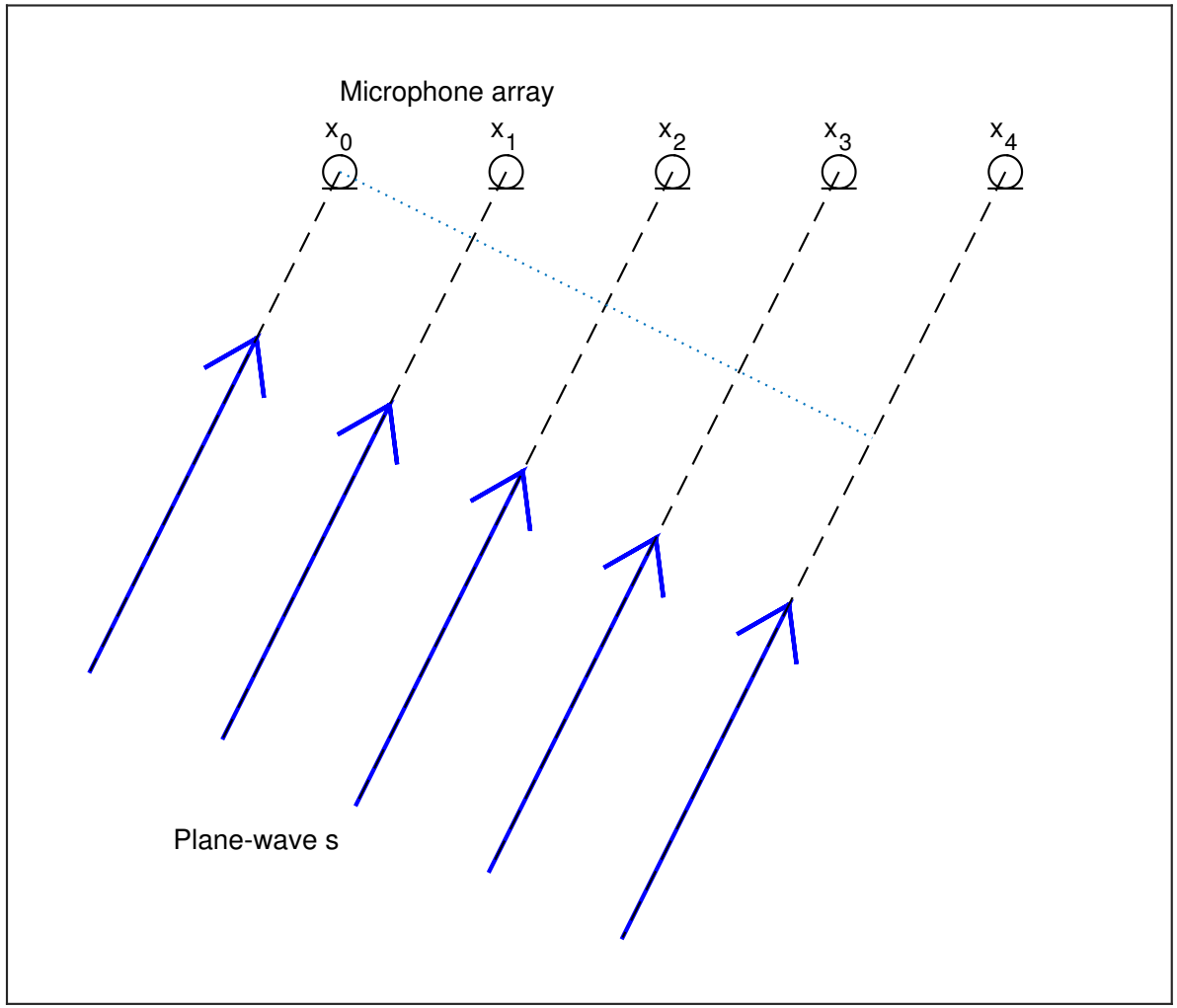
Mikrofonihilan käyttö tilaäänenehostamisessa



Puheensiistaus/-ehostus

Puheteknologian sovelluksia

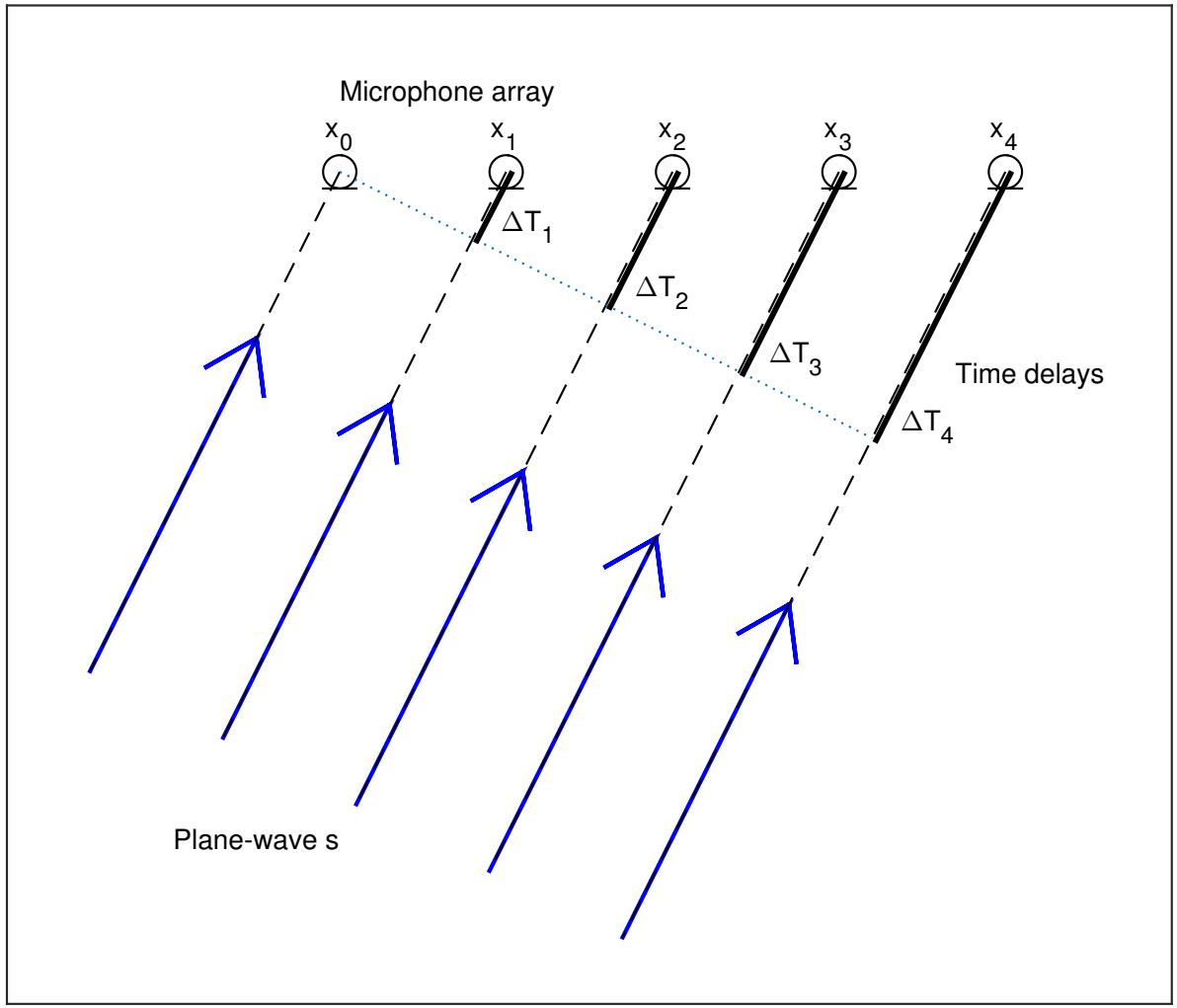
Mikrofonihilan käyttö tilaäänenehostamisessa



Puheensiistaus/-ehostus

Puheteknologian sovelluksia

Mikrofonihilan käyttö tilaäänenehostamisessa



Puheensiistaus/-ehostus

Puheteknologian sovelluksia

Mikrofonihilan käyttö tilaäänenehostamisessa

- Ääni saapuu eri mikrofoneille eri aikaan.
- Viivästyttämällä mikrofonisignaaleja sopivasti saadaan haluttu ääni kohdakkain kaikissa mikrofonisignaaleissa.
- Summaamalla viivästetyt signaalit saadaan haluttu signaali vahvistumaan.
- Häiriösignaalit ovat summassa epäsynkassa (eri vaiheessa), joten ne kumoavat toisensa osittain.

$\Rightarrow$ Haluttu ääni vahvistuu ja häiriöt heikkenevät.

Internet of Things (IoT) ja puhesignaalit

Puheteknologian sovelluksia

- Yhä useampi laite on yhdistetty Internettiin (laitteiden Internet).
- Yhä useampi laite on puheohjattu; televisiot, älykaiuttimet, puhelimet, mikroaaltouunit yms.
- Tämä tarjoaa suuren määrän *potentiaalisia* hyötyjä:
  - Laitteiden ohjaus luonnollisella kielellä ilman hankalia menuja (erit. lapset ja vanhukset).
  - Kaikkien laitteiden ohjaus yhdestä käyttöliittymästä.
  - Etäkäyttöinen, eli useamman metrin päästä (erit. liikuntaesteiset, vanhukset yms.)
- Potentiaaliset haitat ovat myös merkittäviä, erityisesti *yksityisyys*:
  - Saavatko mainostajat kuunnella kotiasi?
  - Saako kämppiksesi selata selainhistoriaasi yhteiskäyttöisissä laitteissa?
  - Voiko rikollinen murtautua ja salakuunnella kotiasi?

Puhekäyttöliittymiä

Puheteknologian sovelluksia

Määritelmä: Puhekäyttöliittymä

- Konetta ohjataan ja käytetään puhumalla
- Joissain tapauksissa kone myös vastaa puhumalla
- “Kone” voi olla hardwarea, softwarea tms.
- Puheentunnistin “ymmärtää” puhetta
- Puhesynteesi vastaa puheella
- “Dialogisysteemit” (dialog system) on se kombinaattori joka kasaa sisääntulevista lauseista merkityksen ja sille sitä vastaavan vastauksen ja toiminnan

Puhekäyttöliittymiä

Puheteknologian sovelluksia

- Puhekomentoja laitteelle
  - Matkapuhelimet, navigaattorit
- Automatisoidut puhelinkeskukset
  - Asiakaspalvelut lentoyhtiössä, puhelinyhtiöissä, taksi jne.
- Sanelusovellukset ja automaattinen tekstitys
  - Esim. terveydenhuollon sanelut
  - YouTube-videoiden automaattinen tekstitys
- Tiedon haku puhelimella tai “älylaitteella”
  - Siri, Alexa, Google Now, MyCroft

Puhekäyttöliittymiä

Puheteknologian sovelluksia

Läheisiä ja asiaan liittyviä teknologioita

- Puhujan tunnistus (esim. smart-home) ja varmennus (esim. pankkipalvelut)
- Emootioiden tunnistus (esim. asiakaspalvelu)
- Äänten tunnistus ja luokittelu (elokuvien indeksointi, turvallisuussovellukset)
- Puheaktiivisuuden tunnistaminen (puhuuko vai ei?)
- Keyword spotting (“Siri!”) ja aiheen tunnistus (“talousuutisissa tänään..”)

Nämä ovat käyttöliittymää tukevia toimintoja jotka antavat lisäinfoa, tai virransäästöön liittyviä toimintoja.

Puheentunnistus – Orientaatio

Puheteknologian sovelluksia

Paritehtävä (3min)

- Miksi automaattinen puheentunnistus on vaikeaa?

Puheentunnistus – Orientaatio

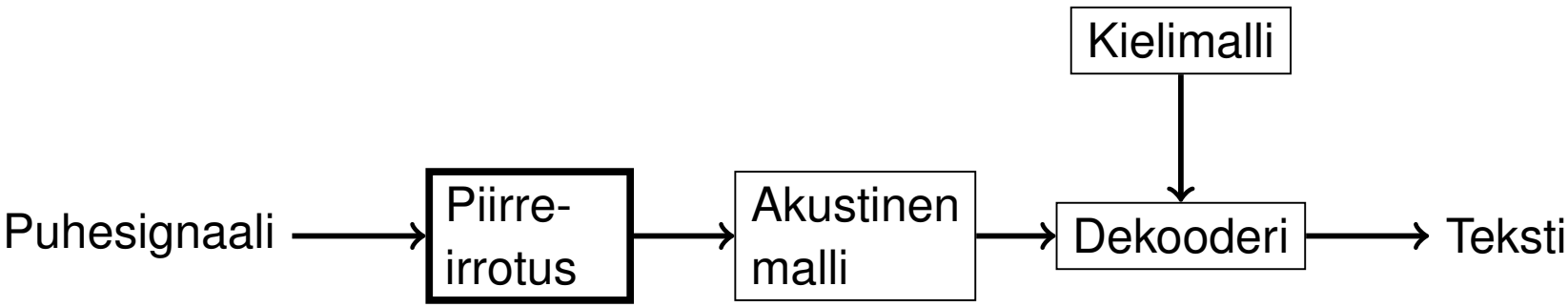
Puheteknologian sovelluksia

- Signaalin laatu
  - Taustamelu: kahvila, auto, liikenne
  - Huoneakustiikka: lähimikrofoni, pöytämikrofoni, huoneen ominaisuudet
  - Laitteisto yms: Mikrofonien laatu, DA-muunnin, pakkaus
- Puhetyyli
  - Erotetut sanat vs jatkuva puhe
  - Rajoitettu vs laaja sanasto
  - Spontaani puhe vs kysmys-vastaus
- Puhuja
  - Puhujariippuvat vs -riippumattomat mallit
  - Adaptaatio

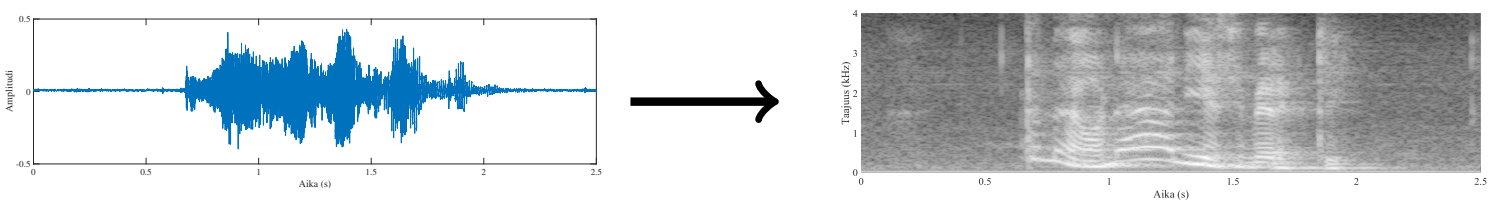
Puheentunnistus

Puheteknologian sovelluksia

- Puheentunnistin muuntaa puhutun äänisignaalin tekstiksi

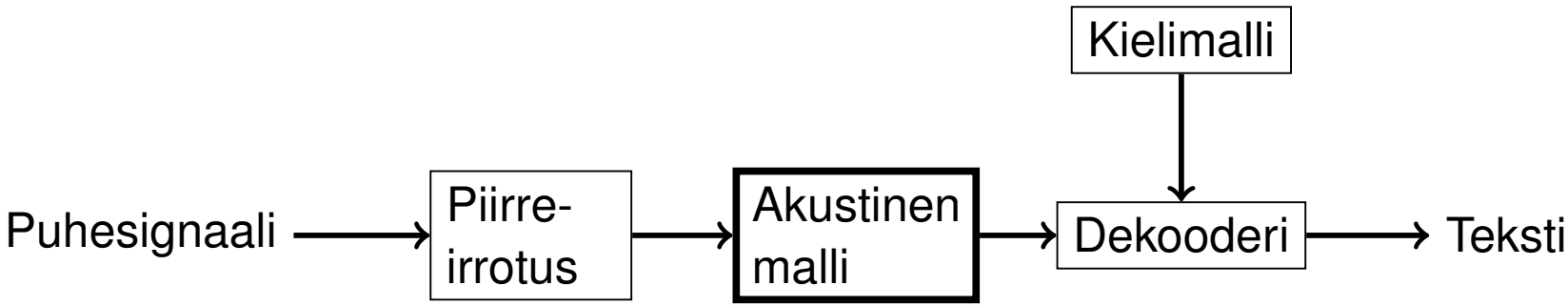


- Piirreirrotuksessa signaalista analysoidaan puhetta kuvaavia mittoja, kuten energiatiheys eri taajuuskaistoilla.
  - Esim. Perkeptuaalisesti (eli kuulon ominaisuuksien mukaan) painotettu spektrogrammi.



Puheentunnistus

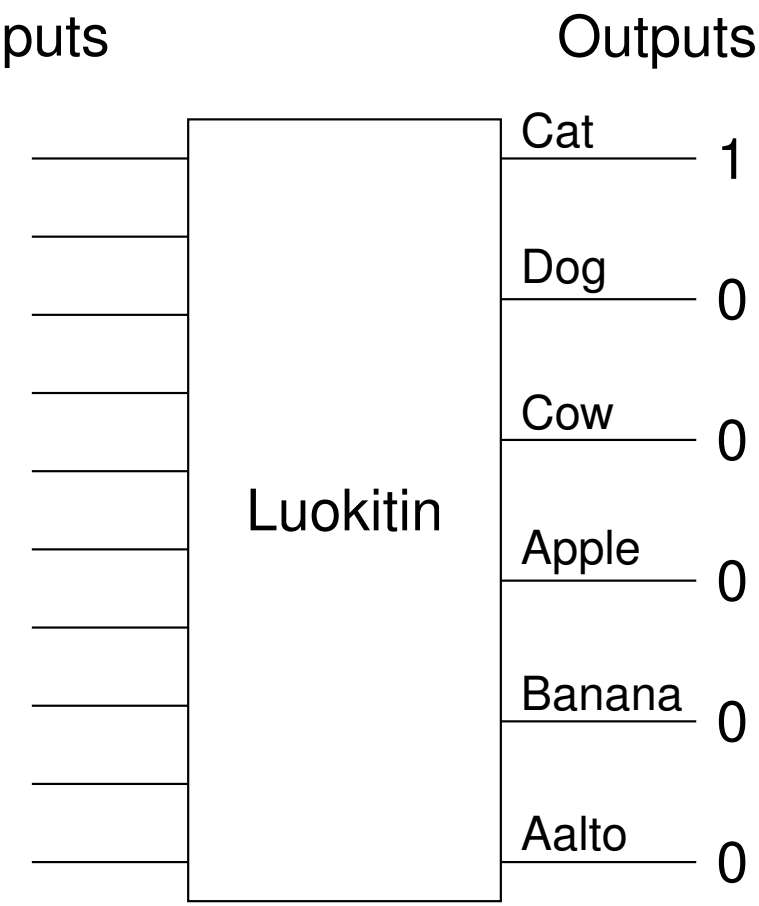
Puheteknologian sovelluksia



- Akustinen malli on tilastollinen malli, eli *luokitin*, joka arvioi millä todennäköisyydellä annettu ääni kuuluu mihinkin foneemiluokkaan (äänneluokkaan).
- Malli on opetettu kymmenistä–tuhansista tunneista ääntä puhuttua kieltä eri puhujilla.

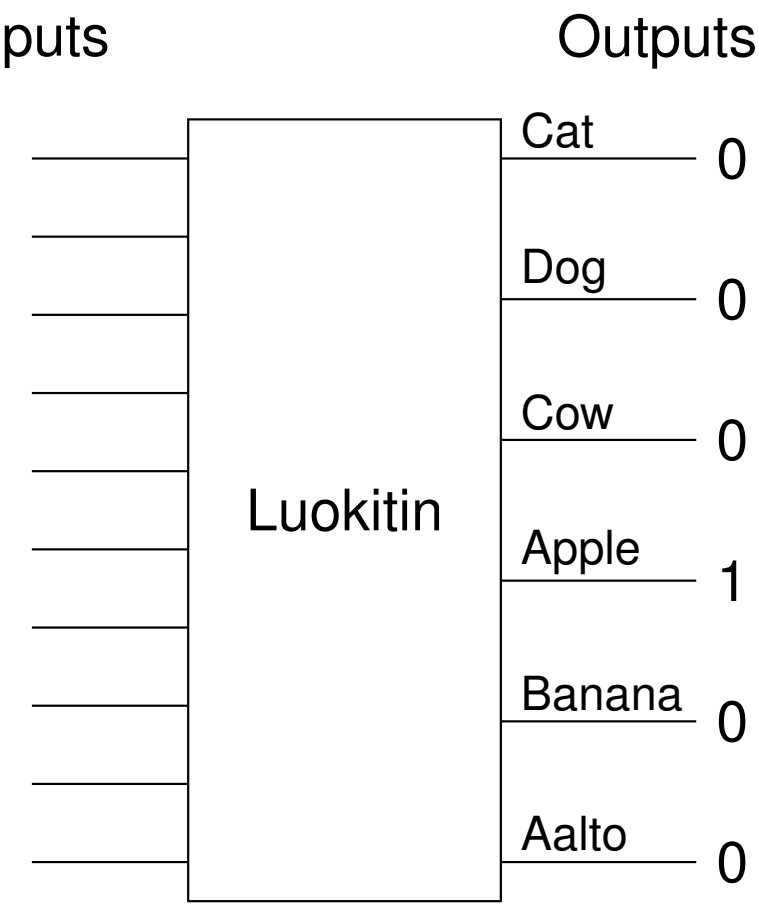
Puheentunnistus – Luokitin

Puheteknologian sovelluksia



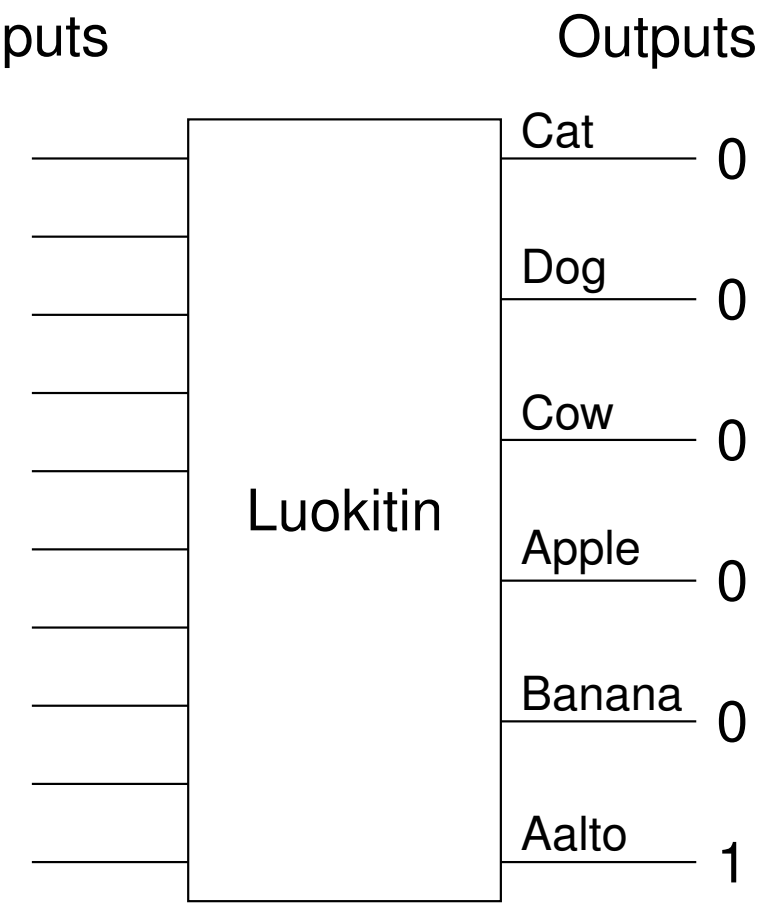
Puheentunnistus – Luokitin

Puheteknologian sovelluksia



Puheentunnistus – Luokitin

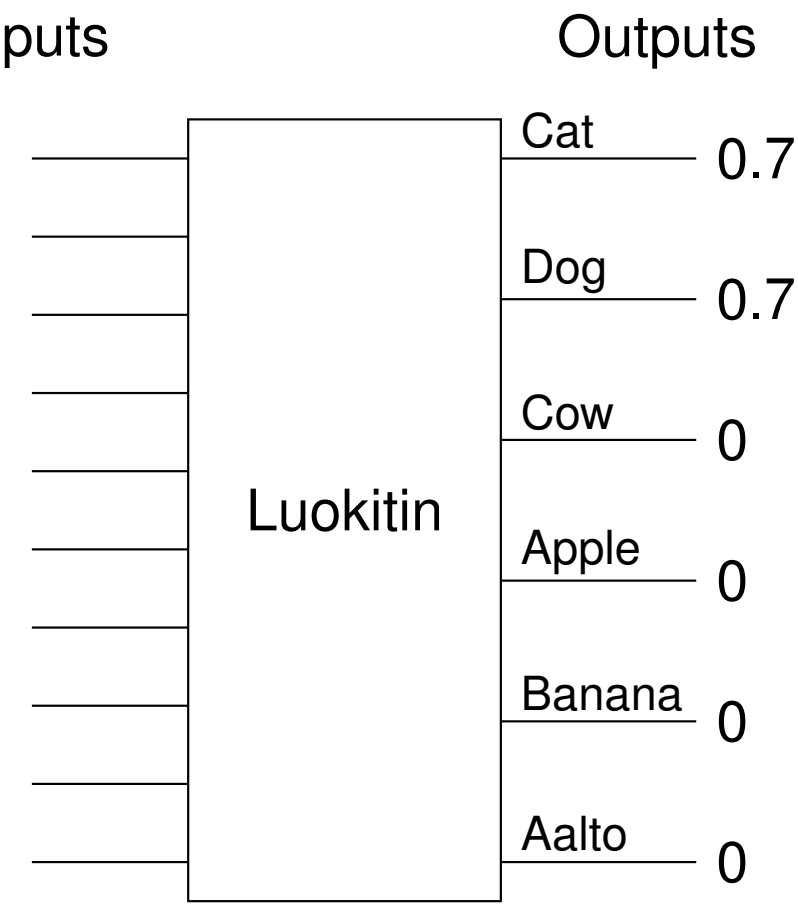
Puheteknologian sovelluksia



Aalto-yliopisto

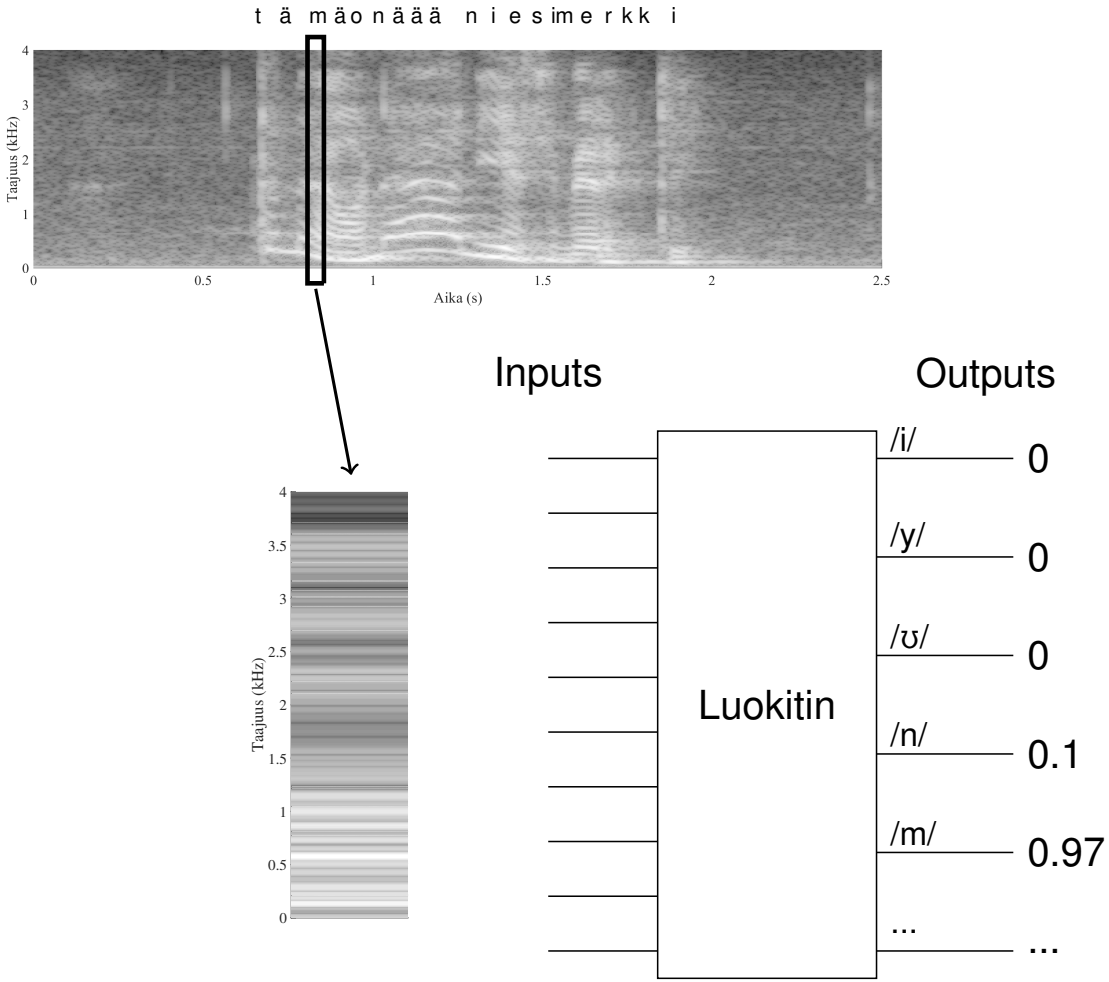
Puheentunnistus – Luokitin

Puheteknologian sovelluksia



Puheentunnistus – Foneemiluokitin

Puheteknologian sovelluksia



Puheentunnistus – Tilastollinen malli / luokitin

Puheteknologian sovelluksia

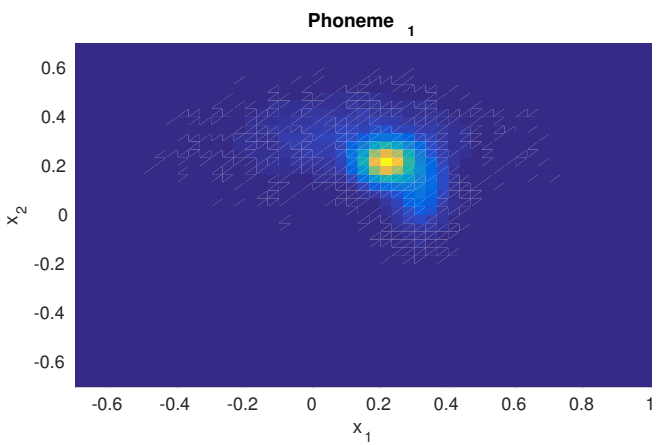
- Yksi perinteinen malli on Gaussin sekoitemalli

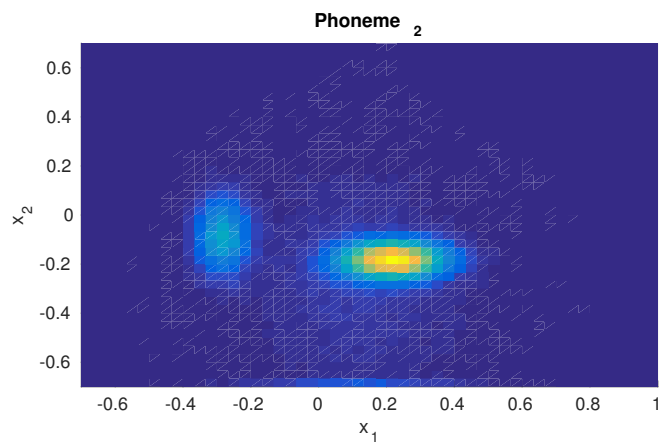
$$p(x, \text{foneemi}_h) = \sum_{k=1}^{M} A_{k,h} \exp\left(-x^H C_{k,h} x\right).$$

missä x on piirrevektori ja $A_{k,h}$:t ovat vakioita ja $C_{k,h}$ kovarianssimatriisi.

- Se on yleistys perinteisestä Gaussisesta todennäköisyydestä

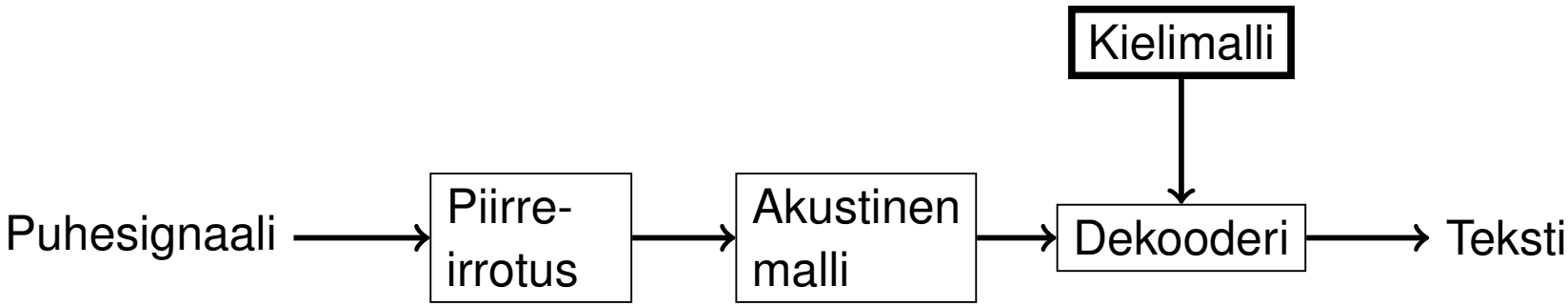
$$p(\xi) = Ce^{-\frac{\xi^2}{2\sigma^2}}.$$





Puheentunnistus – Kielimallit

Puheteknologian sovelluksia



- Kielimalli on tilastollinen malli joka kertoo
  - miten todennäköisesti foneemit seuraavat toisiaan ja
  - miten todennäköisesti sanat seuraavat toisiaan.
- Opetetaan tekstiaineistosta – lehtiä, kirjoja, ym. tavallista tekstiä
- Laajuus luokkaa 10 miljoonaa sanaa

Tehtävä

Mikä sana ennustaa todennäköisimmin sanaa “eat”?

sana	todennäköisyys
I (eat)	0.0038
lunch (eat)	0
to (eat)	0.26

Puheentunnistus – Kielimallit

Puheteknologian sovelluksia

Tehtävä

Mikä sana ennustaa todennäköisimmin sanaa “lunch”?

sana	todennäköisyys
want (lunch)	0.0049
food (lunch)	0
Chinese (lunch)	0.0047

Data from Berkeley restaurant corpus (Jurafsky & Martin, 2000 "Speech and language processing").

	I	want	to	eat	Chinese	food	lunch
I	8	1087	0	13	0	0	0
want	3	0	786	0	6	8	6
to	3	0	10	860	3	0	12
eat	0	0	2	0	19	2	52
Chinese	2	0	0	0	0	120	1
food	19	0	17	0	0	0	0
lunch	4	0	0	0	0	1	0

Uni-gram counts

!	3437
want	1215
to	3256
eat	938
Chinese	213
food	1506
lunch	459

1087 / 3437=.32

3 / 3256 = .00092

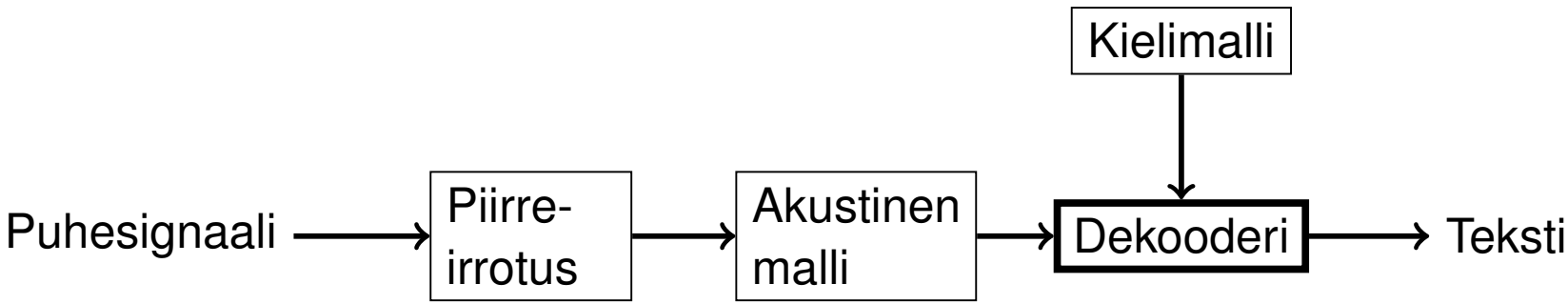
6 / 1215 = .0049

Calculate missing bi-gram probabilities

	I	want	to	eat	Chinese	food	lunch
I	.0023	.32	0	.0038	0	0	0
want	.0025	0	.65	0	.0049	.0066	.0049
to	.00092	0	.0031	.26	.00092	0	.0037
eat	0	0	.0021	0	.020	.0021	.055
Chinese	.0094	0	0	0	0	.056	.0047
food	.013	0	.011	0	0	0	0
lunch	.0087	0	0	0	0	.0022	0

Puheentunnistus – Dekooderi

Puheteknologian sovelluksia



- Dekooderi yhdistää akustisen ja kielimallin
- Valitsee eri tunnistushypoteeseista parhaan

Puheentunnistus – Dekoodaushypoteeseja

Puheteknologian sovelluksia

1. I will tell you would I think in my office
2. I will tell you what I think in my office
3. I will tell you when I think in my office
4. I would sell you would I think in my office
5. I would sell you what I think in my office
6. I would sell you when I think in my office
7. I will tell you would I think in my office
8. I will tell you why I think in my office
9. I will tell you what I think on my office
10. I Wilson you I think on my office

Regenerated from picture by Bryan Pellom

Puhesynteesi

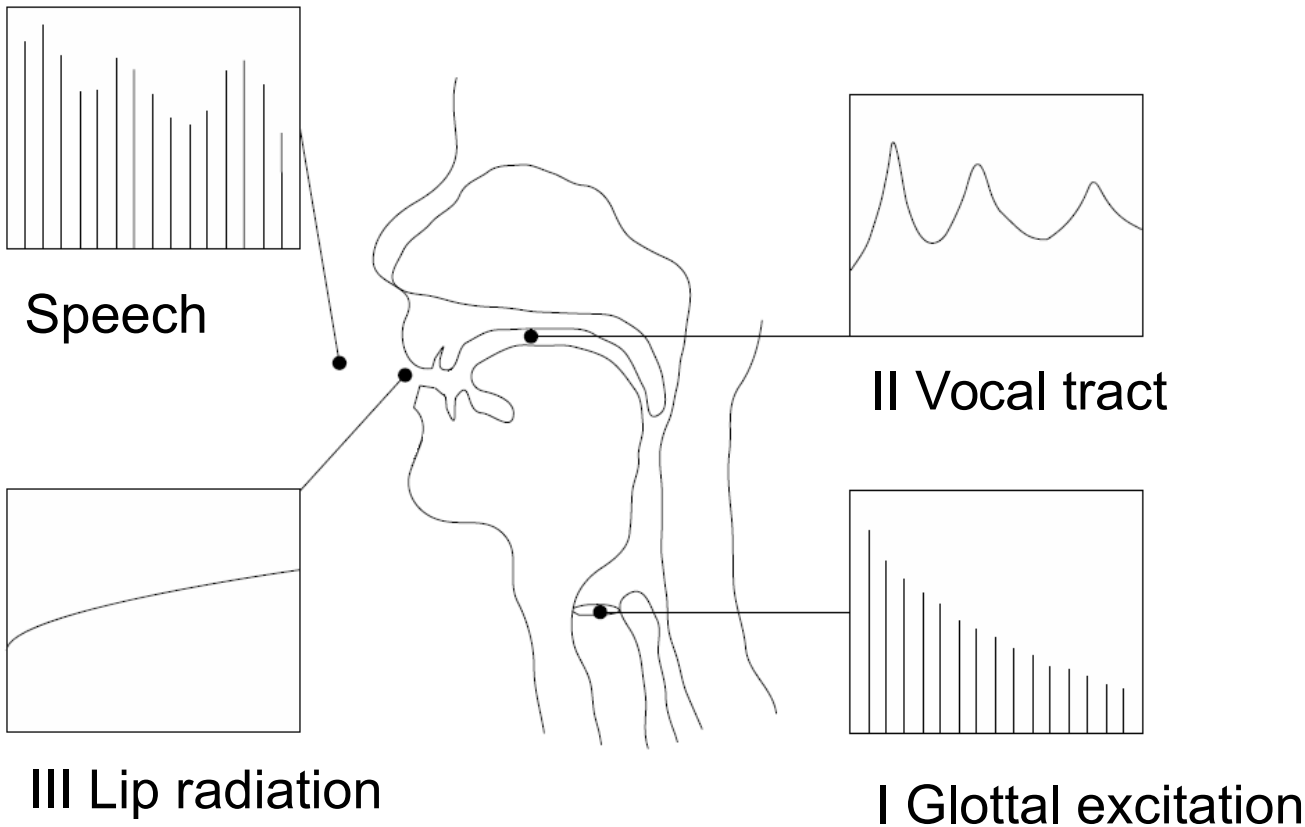
Puheteknologian sovelluksia

- Puhesynteesillä tarkoitetaan puheäänen luomista tietokoneella tms.
  - Input: Tekstiä, Output: Puhetta
- Synteesissä on käytössä kaksi lähestymistapaa
  - Konkatenoiva synteesi – Nauhoitetaan suuri määrä luonnollista puhetta, leikataan siitä irti yksittäisiä äänteitä, ja leikataanliimataan niistä uusia sanoja.
  - Fysikaalinen mallinnus – Mallinetaan puheentuoton fysiologista osaa ja opetetaan sille tekstin ja fysiologian korrelaatio.
- Konkatenoiva synteesi on hyvin, hyvin työlästä, mutta sillä saa yleensä erinomaisen äänenlaadun.
- Synteesi fysikaalisella mallilla on “helppoa” (opetus on raskasta), mutta äänenlaatu ei ole aivan yhtä hyvä.

Puhesynteesi

Puheteknologian sovelluksia

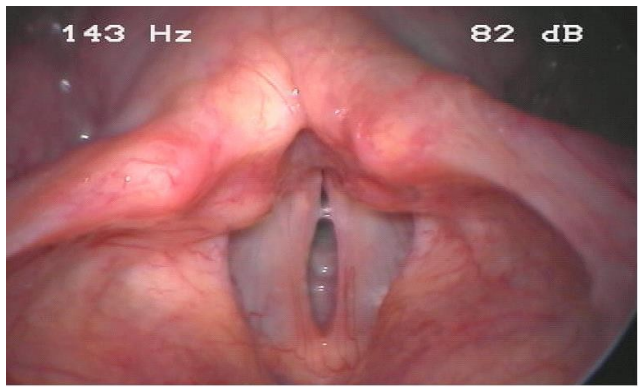
Puhesynteesi puheentuottoa mallintamalla (prof. Paavo Alku)

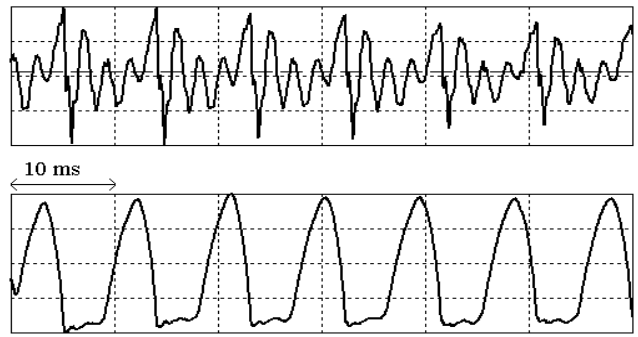


Puhesynteesi

Puheteknologian sovelluksia

Äänihuulimallilla muodostetaan perustaajuus (glottisheräte)

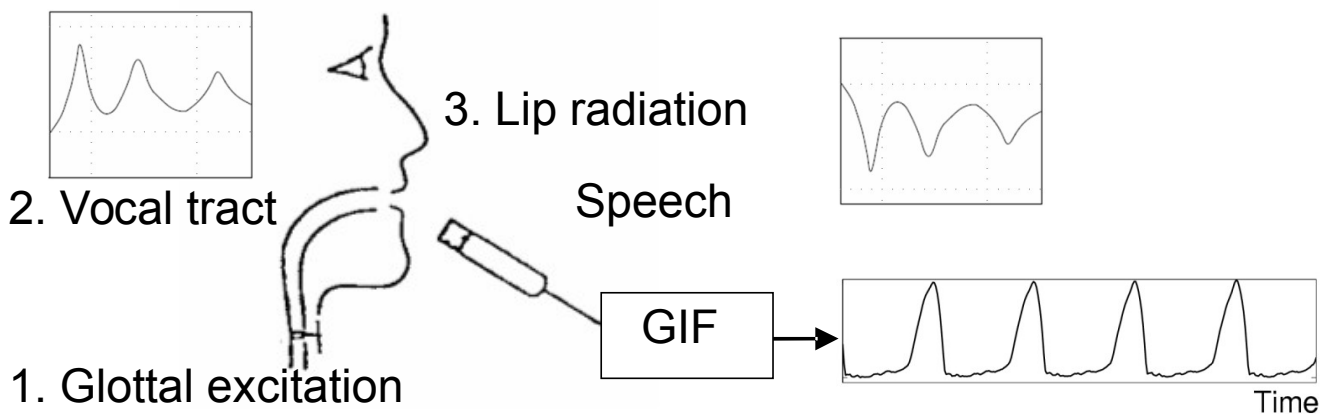




Puhesynteesi

Puheteknologian sovelluksia

- Estimoimalla ääniväylä (vocal tract) ja huulisäteily, voidaan niiden vaikutus kumota (glottal inverse filter, GIF).
- Saadaan aproksimaatio glottisherätteestä.

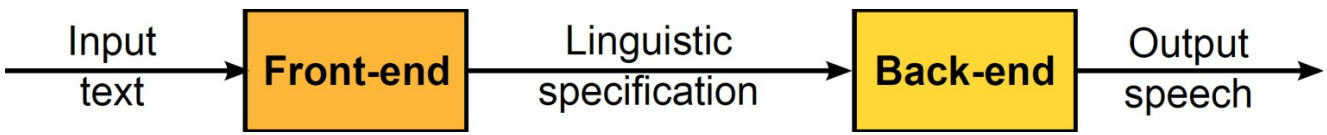


Puhesynteesi

Puheteknologian sovelluksia

Synteesivaihe

- Front-end kääntää tekstin sarjaksi äänteitä
- Back-end tuottaa ääntä fysiologisella mallilla äänteiden perusteella

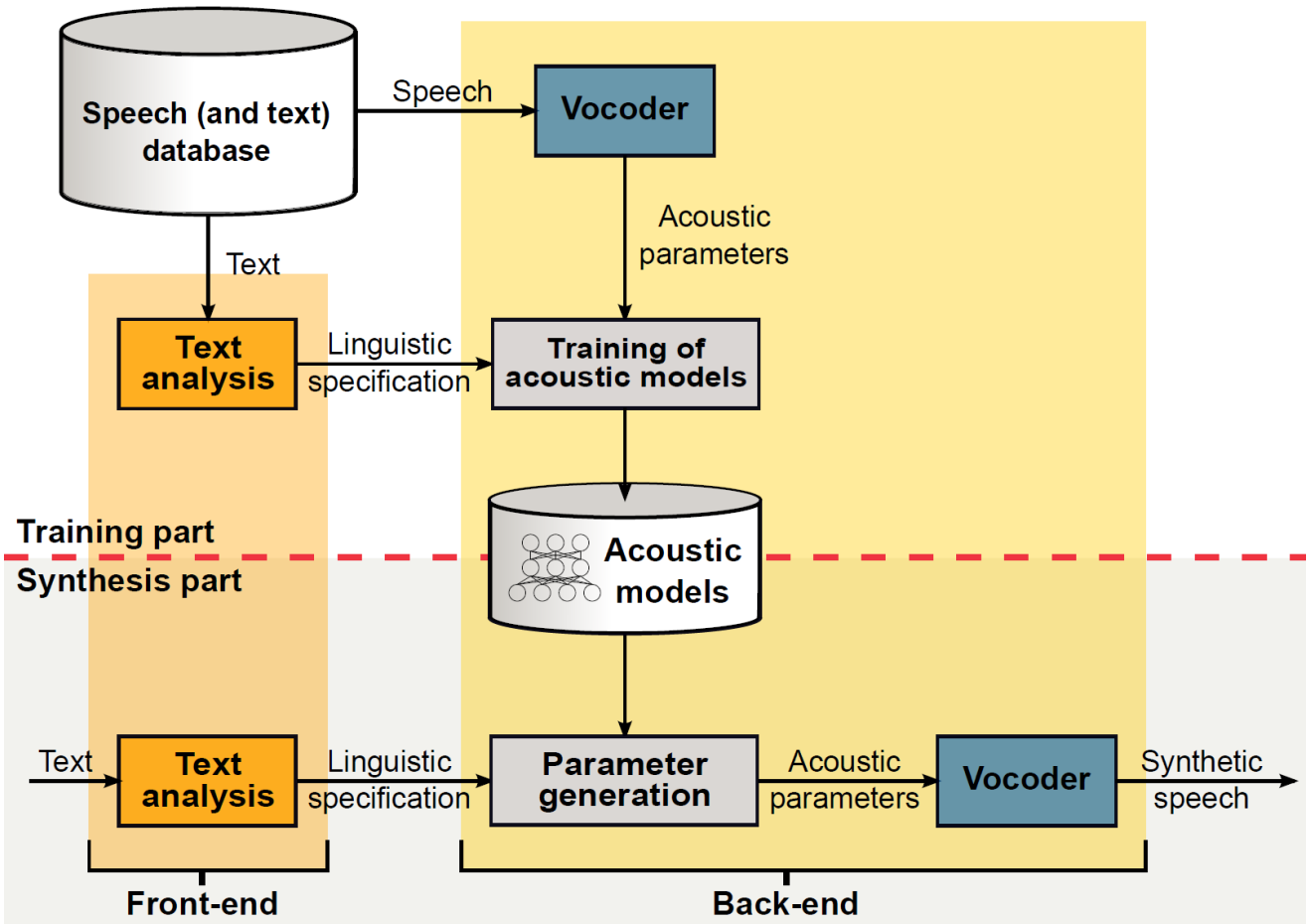


- Front-end vastaa puheentunnistimen dekooderia ja back-end foneemiluokitinta.

Puhesynteesi

Puheteknologian sovelluksia

Opetusvaiheessa opetetaan hermoverkolle äänteiden ja fysiologisten parametrien yhteys



Demoja ryhmien tutkimuksesta

Puheteknologian sovelluksia

- Automaattinen videon sisällön annotaatio
  - `https://youtu.be/wdFA1xAdHGE`
- Automaattinen tulkki joka kääntää käyttäjän puheen suomesta englanniksi
  - EU Emime-projekti
  - `https://youtu.be/wqv7uYAyAQ0`

Sen pituinen se!
Kiitos mielenkiinnosta.