Puheteknologian perusteet
Kurssilla "Informaatioteknologian perusteet"

Tom Bäckström

Kevät 2019

1 Johdanto

Puheääni on ihmisen tärkein kommunikaatiomuoto. Se on monella tapaa ihmisyyttä karakterisoiva asia – se on ehkä suurin ero ihmisen ja eläinten välillä. "Ajattelen, siis olen", sanoi René Descartes viitaten siihen ihmiselle ominaiseen ajatteluun, jota teemme suurelta osin kielen avulla.

Kuva 1: Klassinen vuorovaikutustilanne: naimakaupat.

Siinä missä visuaalinen kommunikaatio on tehokas yksisuuntaiseen tiedonsiirtoon (kts. kuva 2), puhekieli on tehokkaimmillaan vuorovaikutusvälineenä ihmisten välillä (kts. kuva 1). Esimerkiksi kun haet uutta työpaikkaa, sinut todennäköisesti kutsutaan henkilökohtaiseen työhaastatteluun vaikka olet jo toimittanut CV:si kirjallisessa muodossa. Haastattelun vuorovaikutus kun on paljon kirjoitettua viestiä tehokkaampi tapa selvittää oletko juuri sinä oikea henkilö tähän työpaikkaan.

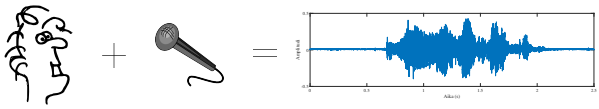
Kuva 2: Yksi kuva kertoo enemmän kuin tuhat sanaa.

Koska puheääni on ihmiselle niin keskeinen asia, ovat puheääneen liittyvät teknologiat myös insinöörille hyvin tärkeitä. Useimmat puheteknologiat liittyvät oleellisesti kännyköihin; Digitaalinen puheensiirto, eli 90-luvun kännykät, oli ensimmäinen modernin puheteknologian voimannäyttö. Nykyään tutkimus käy kuumana henkilökohtaisten assistenttien (personal digital assistant, PDA) kimpussa, kuten Siri, Alexa ja Google, jotka ymmärtävät puhetta ja osaavat vastata kysymyksiin ja suorittaa yksinkertaisia tehtäviä.

Puheteknologialla viitataan kaikkiin niihin teknologioihin jotka jotenkin liittyvät puheääneen. Ala on laaja ja pitää sisällään useampia itsenäisiä osia, kuten

Puheenkoodaus on digitaalista puheäänen pakkausta, missä pyritään mahdollisimman alhaisella datamäärällä per aikayksikkö, mahdollistamaan tiedon siirto tai tallennus siten että ääni voidaan jälkeenpäin toistaa ja vielä siten, että toistetun äänen laatu ja sisältö kuulostaa samalta kuin alkuperäinen.

Puheentunnistus viittaa tekniikoihin joilla pyritään saamaan tietokone ymmärtämään puheääntä ja suorittamaan tehtäviä sen perusteella.



Kuva 3: Puheäänen paineaalto muunnetaan sähköiseksi signaaliksi mikrofonin avulla.

Puhesynteesi viittaa teknologioihin joilla tietokone tuottaa puhetta. Yleensä tämä tarkoittaa sitä että ohjelma ottaa syötteenä tekstiä ja tuottaa sen perusteella ääntä.

Puheensiistaus tai -ehostus on puhesignaalin parantamista. Useimmiten tämä tarkoittaa taustakohinan tai huonekaiun poistamista äänitetystä signaalista, mutta se voi myös tarkoittaa puhesignaalin parannusta siten että vaikka rupinen tai käheä ääni muutetaan sulosointuisaksi tai muuten paremmin ymmärrettäväksi puheääneksi.

2 Puhesignaali

Puheääni on akustinen signaali, eli se muodostuu ilmanpaineen värähtelyistä ilmassa (vrt. kommunikaatioakustiikan luento). Paineaallon voi muuttaa sähköiseksi signaaliksi mikrofonin avulla (kts. kuva 3). Useimmiten se muunnetaan vielä digitaaliseen muotoon AD-muuntimen avulla.

Puhesignaalissa on energiaa lähinnä taajuskaistalla 50–15 000 Hz ja tärkein informaatio on välillä 350–3500 Hz. Tyypillisesti näytteenottotaajuus valitaan siksi väliltä 8–48 kHz. Perinteiset analogipuhelimet siirsivät juuri tämän 300–3300 Hz alueen, mistä johtuen puhe on niissä ymmärrettävää, mutta korkeampien taajuuksien puuttumisen vuoksi niistä puuttuu kirkkaus.

Saadaksemme arvion puheäänen informaatiomäärästä, voimme olettaa että puhesignaalin on näytteistetty 8 kHz:illä ja näytteet esitetään 16 bittisenä. Informaatiomäärä tässä signaalissa on siten 128 kbit/s, tai 440 Mbit/h. Lienee ilmeistä että tässä esitysmuodossa on valtavasti ylimääräistä turhaa tietoa, joka ei ole meidän kannalta oleellista. Voidaksemme arvioida puheäänen informaation määrää, pitää meidän siis lähteä toisesta päästä liikkeelle.

Puheäänen merkitys eli sen informaatio on pääsääntöisesti sen välittämissä sanoissa ja lauseissa. Jos kirjoitetussa kielessä on eri kirjaimia jossain välillä 20-40, kielestä riippuen, voimme karkeasti yleistäen sanoa että yhden kirjaimen voi esittää 5 bitillä. Puhenopeus on useimmiten välillä 5–15 äännettä/s, joten informaatiotiheys puhutussa tekstissä on karkeasti arvioiden noin 75 bit/s. Tämä arvio ei ota huomioon sitä että monet kirjainyhdistelmät ovat paljon todennäköisempiä kuin toiset. Esimerkiksi kirjainyhdistelmä “xjhghgzx” esiintyy äärimmäisen harvoin puhutussa kielessä. Poistamalla nämä “mahdottomat” kirjain- tai äänneyhdistelmät, saadaan paljon pienempi arvio informaatiotiheydestä. Aiheesta lisää puheentunnistus-osiossa.

Samalla on selvää että puheääni pitää sisällään myös paljon muuta informaatiota. Esim. puhelinkeskustelun perusteella voit päätellä puhujasta paljon; puhujan iän, sukupuolen, etnisen ryhmän, koulutustason, terveyteen ja tunnetilaan liittyviä asioita, jne. Puheääntä voi myös kuvata sen äänenkorkeuden

Lingvistinen informaatio
Foneemit (äänteet)
Tavut ja sanat
Lauseet yms.
Kieli, murre ja aksentti
Paralingvistinen informaatio (muu kuin lingivistinen)
Tila: Henkilöllisyys, terveys, tunnetila, ikä, sukupuli
Tyyli ja oheisviestintä: Puhetyyli, sosiaalinen asema ja konteksi, asioiden painotus
Fysiologiset piirteet: Kielen, huulten, leuan yms. ja niitä ohjaavien lihasten koko, kireys ja asento, sekä ilmavirran määrä
Akustiset piirteet: Äänenkorkeus, kireys, huokoisuus, puhenopus

Taulukko 1: Muutamia puheäänen informaatiotyyppejä.

ja esimerkiksi kireyden mukaan tai fysiologisten piirteiden kuten kielen ja huulten liikkeiden mukaan. Puheääni siis kantaa mukanaan tietoa näistä kaikista piirteistä.

Kuten näimme, *lingvistinen* eli kieleen liittyvä informaatio on noin 75 bit/s, mutta ei ole lainkaan selvää paljonko *paralingvististä* informaatiota (muu kuin lingivistinen) puheääni kantaa mukanaan. Lisäksi suuri osa paralingvistisestä tiedosta on tila-informaatioita, eli se ei muutu ajan yli, joten on vaikea määrittää mitä informaatiotiheys tarkoittaa tässä yhteydessä. Intuitivisesti on kuitenkin selvää että puheen tekstisisältö on sen tärkein sisältö, joten paralingvistinen informaatiosisältö on tuskin lingvististä suurempaa. Tämä käsienheilutteluargumentti johtaa arvioon että puheäänessä on noin 150 bit/s informaatioita. Ero digitaaliseen signaalin on hurja; informaatiotiheys digitoidussa äänessä on liki 1000-kertainen! Eräs puheteknologioiden keskeisistä tavoitteista on siten tuon olennaisen informaation, 150 bit/s kaivaminen esiin informaatiotulvasta.

2.1 Hieman fysiologiaa

Äänentuotto alkaa aivojen päätöksellä tuottaa ääntä (aie). Tätä seuraa lauserakenteen ja sanojen muodostus ja valinta, sekä niiden hermoimpulssien lähettäminen jota tarvitaan ohjaamaan lihaksia äänen muodostamiseksi. Aivot jatkavat äänen tuoton ohjausta koko äänentuoton ajan, seuraamalla aistimuksia kuulosta ja lihaksista, sekä hienosäätämällä lihasten ohjausta näiden aistimusten perusteella.

Varsinainen fysiologinen äänentuotto alkaa keuhkoista joiden lihaksia puristamalla aikaansaadaan keuhkoihin ylipaine, jonka seurauksena ilma alkaa virrata suun ja nenän kautta ulos. Ilmavirtaus ei sinällään vielä tuota värähtelyjä ilmassa. Ääni muodostuu sitten kahden eri prosessin avulla:

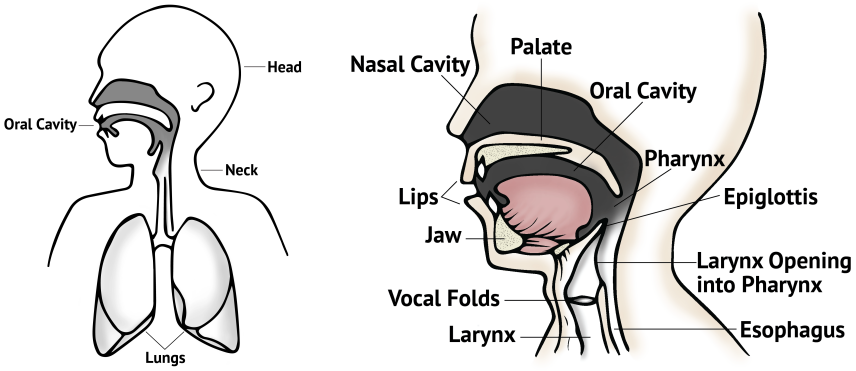
Soinnilliset äänteet muodostuvat kun äänihuulia kiristetään sopivasti, siten että ne alkavat värähtelemään ilmavirrassa (kts. kuva 5). Tyypillisiä soinnillisia äänteitä ovat vokaalit.

Äänihuulten toiminta on havainnollistettu seuraavissa videoissa:

`https://www.youtube.com/watch?v=mJedwz_r2Pc`

`https://www.youtube.com/watch?v=W-nS9fgs7Ro`

(a) (b)



English	Suomi	English	Suomi
nasal cavity	nenäväylä	glottis	äänirako
palate	kitalaki	vocal folds	äänihuulet
oral cavity	suuväylä	larynx	kurkunpää
lips	huulet	esophagus	ruokatorvi
tongue	kieli	vocal tract	ääniväylä
jaw	leuka	vocal fold	äänihuuli
pharynx	nielu	trachea	henkitorvi
epiglottis	kurkunkansi	epiglottis	kurkunkansi

Kuva 4: Äänentuoton anatomiset osat (kirjasta [2]).

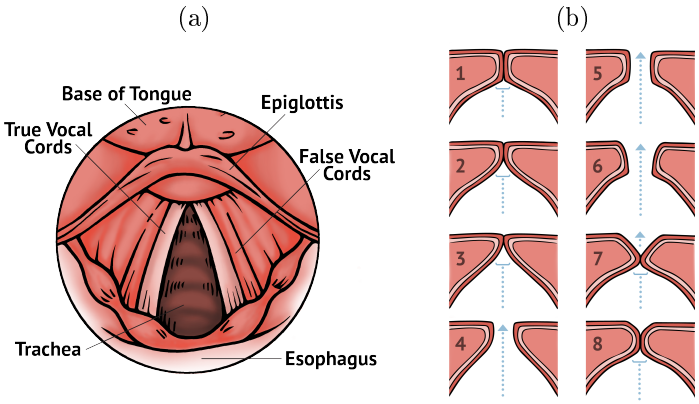
Soinnittomat äänteet muodostuvat ääniväylän ahtaumissa, johon ilmavirta aiheuttaa turbulenssia (epälineaarinen virtausilmiö). Esimerkkejä soinnittomista äänteistä ovat /p/, /s/ ja /h/, joissa turbulenssi muodostetaan vastaavasti huulien väliin, kielen ja hampaiden väliin, tai viemällä kurkunpäätä taaksepäin.

Viimeisenä fysiologisena osana on ääniväylä, eli maalikkotermein se putkisto jota myöden ilma virtaa äänihuulilta suun ja nenän kautta ulos. Niin kuin mikä tahansa muukin putki kuten trumpetti tai huilu, myös ääniväylä resonoi eri taajuuksilla eri tavoin. Resonassit voimistavat joitakin taajuuksia ja heikentävät toisia, samoin kuin vanhan ajan hifi-laitteiden ekvalisaattorit. Näillä ääniväylän resonansseilla, joita *formanteiksi* kutsutaan, on hyvin suuri merkitys äänteiden muodostamisessa. Muokkaamalla ääniväylän muotoa, voi puhuja muutta resonassien, eli formanttien taajuutta.

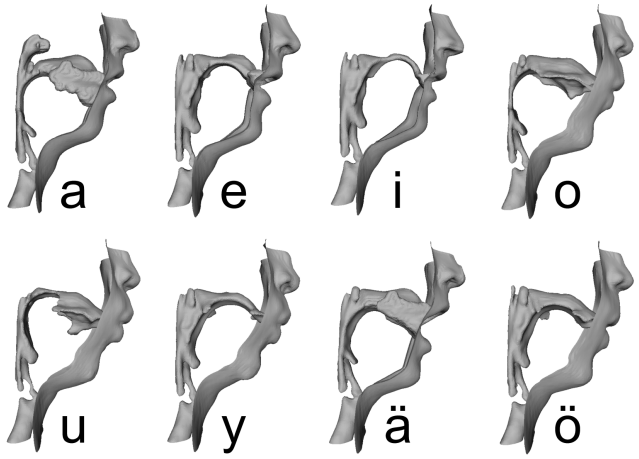
Formantit ovat merkittäviä varsinkin vokaalien kannalta. Kaksi taajuudeltaan alinta formanttia, eli ne kaksi formanttia jotka sattuvat taajuusalueelle 300–1500 Hz, määrittävät vokaalit. Toisin sanoen, se mihin vokaaliin soinnillinen äänne luokitellaan, voidaan määrittää yksinomaan formanttien taajuuden mukaan (kts. kuva 6).

2.2 Hieman fonetiikkaa

Puhetta voidaan esittää kirjoitetussa muodossa tekstinä. Kirjoitusmuoto, joka on jokaisella kielellä omansa, välittää yksiselitteisesti sen viestin jota puheella



Kuva 5: (a) Äänihuulet kuvattuna ylhäältä, sekä (b) äänihuulten jaksollinen liike ilmavirrassa kuvattuna poikkileikkauksena edestäpäin (kirjasta [2]).



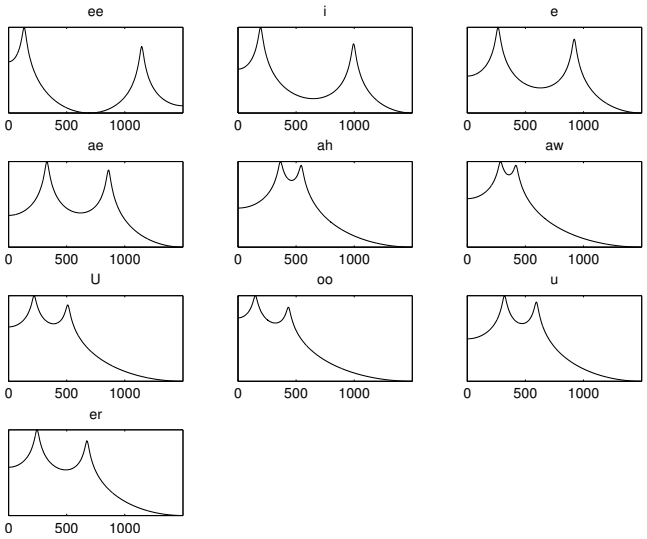
Kuva 6: Ääntöväylän muodot suomen kielen eri äänteillä. ©Jarmo Malinen

oli tarkoitus ilmaista. Vaikka puheen merkitys saakin näin yksiselitteisen tallennusmuodon, ei se ole lainkaan yhtä yksiselitteinen ääntämyksen suhteen. Eri puhujilla on erillaisia tapoja ääntää, eri murteita ja aksentteja, niin että he kuulostavat hyvinkin erilaiselta vaikka he ääntävät samoja sanoja.

Siinä missä kirjoitetun kielen pienin yksikkö on kirjain, niin puhutussa kielessä pienin yksikkö on *foneemi*. Se määritellään pienimmäksi äänen osaseksi jolla on merkitys. Kääntäen, vaihtamalla foneemia, voidaan sanan merkitystä muuttaa, kuten kissa/kassa, pieni/sieni jne.

Jotta eri ääntämyksiä ja foneemeja voitaisiin tutkia ja niistä voitaisiin keskustella, tarvitaan niitä kuvaamaan yhteinen aakkosto. International Phonetic Alphabet, tai tuttavallisesti IPA, on foneettinen aakkosto, eli se kuvaa luonnollisen kielen äänteitä ja foneemeja.

Oheisista taulukoista 2 ja 3 löytyy joitain esimerkkejä IPA merkeistä. Tarkempi kuvaus löytyy Wikipediasta:
`https://fi.wikipedia.org/wiki/Kansainv%C3%A4linen_foneettinen_aakkosto`.



Kuva 7: Ääntöväylän formantit (resonanssit) taajuusakselilla englannin kielen eri vokaaleille.

Taulukko 2: Esimerkkejä vokaalien IPA merkeistä.

IPA	Esimerkki	IPA	Esimerkki
i	cit**y**, s**ee**, m**ea**t	ɤ	*German:* m**ü**ssen
y	*German:* **ü**ber, R**ü**be	o	*German:* **O**fen, R**o**man
ɨ	ros**e**'s	ɛ	b**e**d
ʉ	r**u**de	œ	*German:* H**ö**lle, g**ö**ttlich
ɯ	*Irish:* c**ao**l	ɜ	b**i**rd
u	**ou**gh, y**ou**, thr**ew**	ɞ	*Irish English:* b**u**t
I	s**i**t	ʌ	r**u**n, w**o**n, fl**oo**d
Y	*German:* f**ü**llt	ɔ	l**a**w, c**au**ght, **a**ll
ʊ	p**u**t, h**oo**d	æ	c**a**t, b**a**d
e	*German:* G**e**nom, M**e**than	ɐ	*German:* od**er**
ø	*French:* p**eu**	a	h**a**t
ə	**a**bout, **a**ren**a**	ɶ	*Swedish:* h**ö**rt
ɘ	*Dutch:* **i**k	ɑ	f**a**ther
ɵ	*Australian English:* b**i**rd	ɒ	n**o**t, l**o**ng, t**a**lk

Tyypillisesti IPAa käytetään esimerkiksi sanakirjoissa, missä sanat esitetään usein sekä niiden tavallisessa kirjoitusasussa että niiden ääntämisohje IPA aakkosilla:

sana	ääntäminen (IPA)
taajama	/'tɑː.jɑ.mɑ/
teekkari	/'teːkːɑri/
mörkö	/'mørkø/

Kuten tässä, äänteen alkaminen ja loppuminen ilmaistaan vinoviivoilla '/·/' tai hakasuluilla '[·]'.

Taulukko 3: Esimerkkejä konsonanttien IPA merkeistä.

IPA	Esimerkki	IPA	Esimerkki
b	**b**uy, ca**b**	θ	**th**igh, ma**th**
d	**d**ye, ca**d**, **d**o	p	**p**ie, s**p**y, ca**p**
ð	**th**y, brea**th**e, fa**th**er	r	**r**ye, t**r**y, ve**r**y (trill)
ʤ	**g**iant, bad**g**e, **j**am	ɹ	**r**ye, t**r**y, ve**r**y (approximant)
f	**ph**i, ca**ff**, **f**an	s	**s**igh, ma**ss**
g	**g**uy, ba**g**	ʃ	**sh**y, ca**sh**, emo**t**ion
h	**h**igh, a**h**ead	t	**t**ie, s**t**y, ca**t**, a**t**om
j	**y**es, **y**acht	tʃ	**Ch**ina, ca**tch**
k	s**k**y, **c**ra**ck**	v	**v**ie, ha**v**e
l	**l**ie, s**l**y, ga**l**	w	**w**ye, s**w**ine
m	**m**y, s**m**ile, ca**m**	z	**z**oo, ha**s**
n	**n**igh, s**n**ide, ca**n**	ʒ	equa**t**ion, plea**s**ure, vi**s**ion, bei**g**e
ŋ	sa**ng**, si**n**k, si**ng**er		

2.3 Puheäänen mallinnus

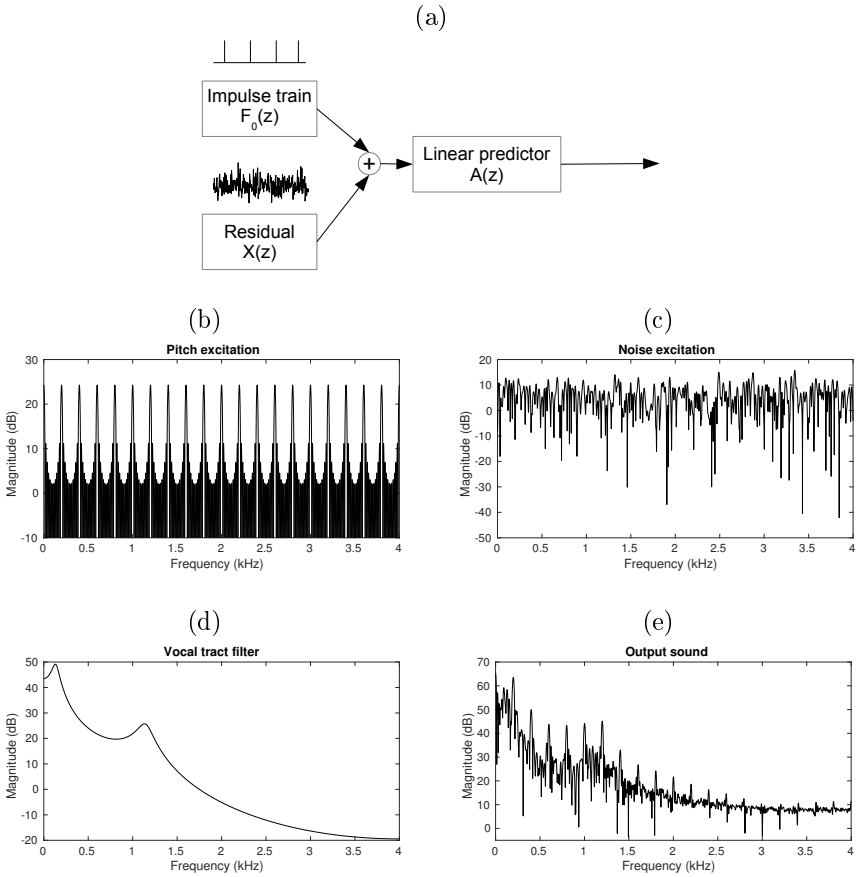
Jotta puheääntä voitaisiin kätevästi käsitellä, pitää signaalille etsiä sellainen esitysmuoto, eli parametrisaatio, joka on siinä mielessä tehokas että kaikki signaalin tärkeät ominaisuudet ovat helposti muokattavissa ja parametrien merkitys on helposti ymmärrettävissä. Hyvä lähtökohta on ihmisen fysiologia. Jos ääntä kuvataan mallilla joka on rakenteeltaan samantapainen kuin ihmisen fysiologia, voimme muokata sen parametreja (kuten vaikka leuan asentoa) siten että voimme jo ennalta arvata miten se muuttaa ääntä.

Klassinen fysiologiaan pohjautuva puheäänen malli on lähde-suodin malli, jossa soinnillisia äänteitä kuvataan impulssijonolla, soinnittomia äänteitä kohinalla, ja ääniväylää suodattimella (kts. kuva 8). Tässä siis impulssijono ja kohina ovat äänilähteitä ja ääniväylä on suodin, ja oletamme että ne toimivat toisistaan riippumattomasti. Impulssijono viittaa signaaliin joka on enimmäkseen nollaa, mutta jossa on impulsseja ($\approx$ ykkösiä) perustaajuutta vastaavin välein. Impulssijonon spektrillä on nk. kampa-rakenne (kts. kuva 8(b)), jossa näkyy perustaajuus taajuudella F_0 =200 Hz sekä sen monikerrat kF_0, kun $k \in \{1, 2, 3, \dots\}$. Ulos tuleva ääni (kuva 8(e)) on siis kahden lähteen summa, kerrottuna suotimen muodolla.

Tätä lähden-suodin mallia on käytetty erityisesti puheenkoodauksessa, jossa se on CELP-metodin perustana.

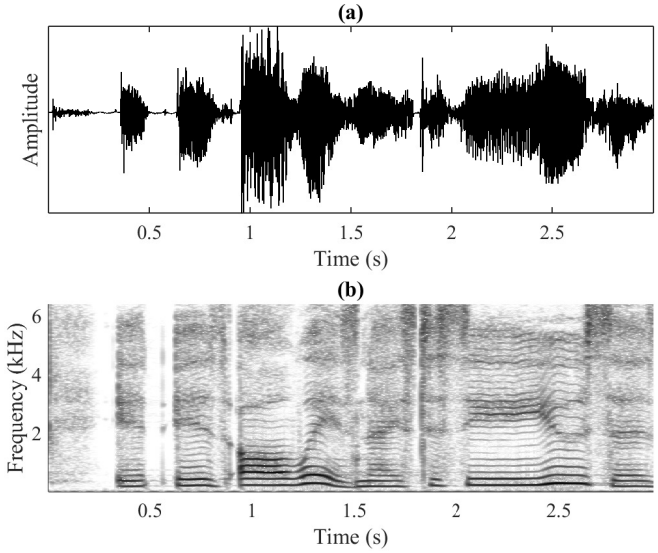
Lähde-suodin malli on tyyppiesimerkki aika-tason mallinnuksesta, missä kaikki komponentit voidaan helposti ilmaista aika-alueen operaatioina; Molemmat lähteet voidaan muodostaa aika-signaaleina ja suodin voidaan toteuttaa digtaalisena IIR-suotimena. Vaihtoehtoinen lähestymistapa on käyttää mallinnusta aika-taajuus tasossa, missä signaalista otetaan lyhyitä pätkiä ja jokaisesta pätkästä analysoidaan sen spektri. Tässä esitysmuodossa voidaan tarkkailla taajuuskomponenttien kehitystä ajan yli.

Esimerkiksi, kuvassa 8(e) on kuva puheäänen spektristä yhtenä ajan hetkenä, mistä nähdään hyvin sekä perustaajuuden kamparakenne sekä formanttien sijainnit (n. 250 Hz ja 1200 Hz). Tämä spektri voidaan kääntää pystysuoraksi vektoriksi, siten että tummat värit kuvaavat korkeita pisteitä ja vaaleat mata-

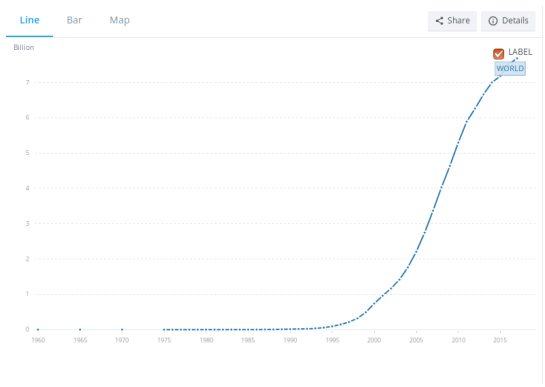


Kuva 8: (a) Lähde-suodinmalli puheentuotossa, (b) perustaajuus-lähde, (c) kohina-lähde, (d) ääniväylä-suodin ja (e) ulostuleva ääni.

lia. Piirtämällä perättäisiä vektoreita vierekkäin nähdään miten spektri muuttuu ajan yli (kts. kuva 9). Tällaista kuvaa kutsutaan spektrogrammiksi. Siinä näkyvät edelleen sekä kamparakenne (tiheät tummat horisontaaliset viivat) että formantit (tumemmat alueet). Kamparakenne on luonnollisesti näkyvissä vain soinnillisten äänien aikana, kun soinnittomat äänteet ovat tasaisenharmaita.



Kuva 9: Puheääni (a) aikasignaalina ja (b) spektrogrammina.

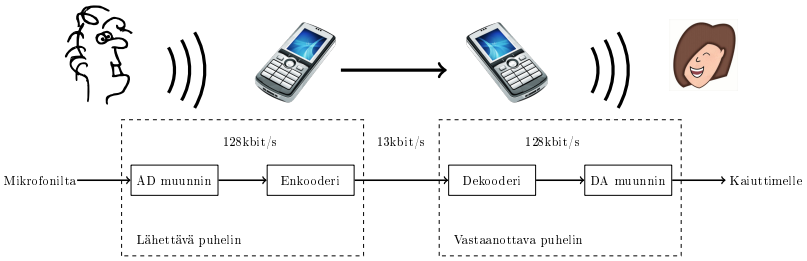


Kuva 10: Mobiililiittymiä per 100 henkilöä maailmassa. Lähde Maailmanpankki.

3 Puheteknologian sovellus: Puheenkoodaus

Digitaalinen mobiilivallankumous alkoi 90-luvulla kun GSM-puhelimet tulivat markkinoille. Suomalainen Radiolinja avasi GSM-verkkonsa asiakkaille ensimmäisenä maailmassa, 1.7.1991. Kymmenen vuotta myöhemmin puhelinliittymiä oli jo 1.5 miljardia ja tänään aktiivisia liittymiä on jo enemmän kuin ihmisiä maapallolla. Toki tämä ei tarkoita sitä että kaikilla ihmisillä olisi puhelinliittymä, koska monilla on useampia liittymiä.

Puheenkoodauksella viitataan äänen digitaaliseen pakkaukseen sen siirtoa ja tallennusta varten. Sovelluksen tavoitteet ovat lähtökohtaisesti ristiriitaisia tai vastakkaisia. Toisaalta haluaisimme käyttää mahdollisimman vähän resursseja – siirtolinjan kaistanleveyttä (bit/s), CPU-kapasiteettia ja RAM/ROM-muistia – mutta samalla haluaisimme että rekonstruoitu puhe olisi mahdollisimman hy-



Kuva 11: Havainnoillistava kuva puheenkoodauksesta.

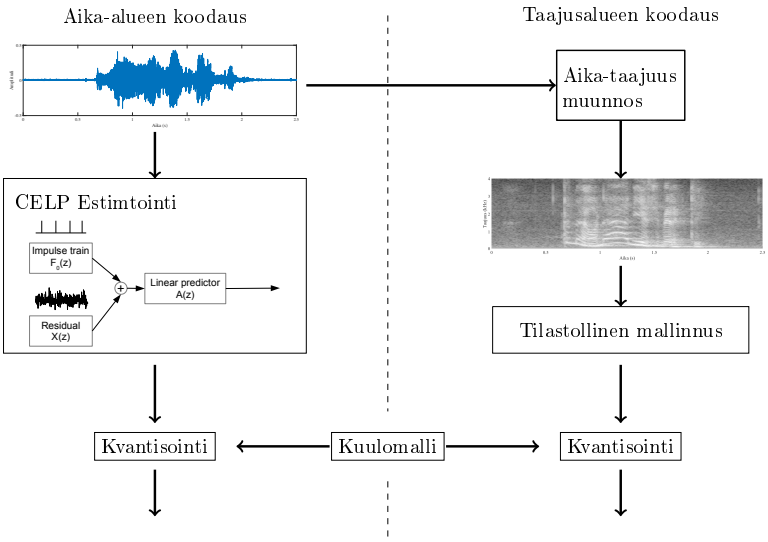
vänlaatuista (ei kohinaa eikä säröä) ja että se voitaisiin toistaa mahdollisimman alhaisella viiveellä.

Tyypillisesti puheenkoodausessa ääni pakataan 10%:iin alkuperäisestä, eli jos 8 kHz:n näytteenottotaajuudella pakkaamaton signaali on 128 kbit/s, niin pakattu signaali on noin 13 kbit/s. Tämä pakkaus on häviöllistä pakkausta, eli tarkoituksena ei ole lainkaan saada täsmälleen samaa signaalia vastaanottajapäässä, vaan on täysin riittävää että rekonstruoitu signaali *kuulostaa* samalta. Tehokkaan pakkauksen kehittelyssä on siksi hyvin tärkeä tietää miten ja mitä ihminen kuulee ääniä. Tärkeimpänä yksittäisenä kuulon piirteenä käytetään hyväksi sitä seikkaa että äänekäs ääni peittää alleen muut lähellä taajuudessa (tai ajassa) olevat äänet. Esimerkiksi jos puheäänessä on voimakas komponentti taajuudella 500 Hz:iä, niin se peittää alleen heikon äänen taajuudella 480 Hz:iä. Näin ollen 500 Hz:in läheisyydessä voidaan kaikki äänet kvantisoida pienemmällä tarkkuudella kuin 1500 Hz:in läheisyydessä, jos siellä ei ole voimakkaita ääniä.

Kuuloa mallintavia menetelmiä kutsutaan perkeptuaalisiksi malleiksi (perceptual model) ja ne ovat niin puheen kuin audionkin koodauksessa erittäin tärkeitä. Perkeptuaalisten mallien käyttö voi hyvinkin parantaa häviöllisen pakkauksen tehokkuuta vaikkapa kertoimella 5 (esim. 50%:sta 10%:iin).

Klassinen lähestymistapa puheenkoodaukseen on lähde-suodin algoritmi (kts. aiempi kappale), johon perustuvat metodit tunnetaan nimellä *Code-Excited Linear Prediction (CELP)* [2]. Kaikki modernit puheenkoodausstandardit, kuten GSM, AMR-WB ja EVS perustuvat CELP-algoritmiin ja sen keksimisellä tai sen kehittämisen onnistumisella oli tärkeä rooli mobiilin vallankumouksen mahdollistajana. CELP-metodissa ääniväylää mallinnetaan IIR-suotimella (=lineaarinen ennustin = linear prediction) jolle annetaan syötteenä (excitation) kohinaa ja impulssijono. Algoritmi on hyvin tehokas kun tiedonsiirtonopeus on 13 kbit/s:n suuruusluokkaa, mutta se ei helposti taivu korkeammille eikä matalemmille nopeuksille. Sitä on myös vaikea laajentaa monikanava-signaaleihin, kuten stereo-äänelle, eikä se myöskään sovellu musiikille.

Uusimmissa puhe- ja audiokoodausstandardeissa on tästä syystä CELP-algoritmin rinnalla myös toinen, aikataajusaluella toimiva koodekki. Näissä koodekeissa kvantisoidaan signaalin spektriä – tavallaan ne kvantisoivat siis signaalin spektrogrammia. Musiikin ystäville tutut MP3-, AAC- ja Ogg-koodekit toimivat samalla periaatteella. Taajuusalueen koodekit ovat CELP:iä helpompia toteuttaa ja ne ovat tehokkaampia (sekä laadultaan parempia että laskennaltaan kevyempiä) korkeilla siirtonopeuksilla. Matalilla tiedonsiirtonopeuksilla taajuusalueen koodekit kuitenkin häviävät äänenlaadussa CELP-metodiin pe-



Kuva 12: Aika- ja taajusalueen koodaus.

rustuville menetelmille.

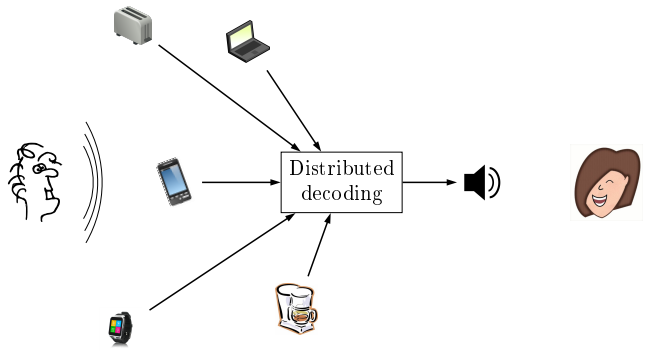
Puheenkoodauksen tutkimus on murrosvaiheessa. Olemme juuri saaneet valmiiksi uuden kansainvälisen standardin, 3GPP Enhanced Voice Services (EVS) vuonna 2014, kaikkien alan suurimpien yritysten ja yliopistojen yhteistyönä. Tämä uusi standardi on jo kaikissa suurimpien valmistajien lippulaivamalleissa (esimerkiksi Qualcomin Snapdragon 820 piirisarjaa käyttävät mallit) ja odotettavissa on että operaattorit ottavat sen pian käyttöön. Tässä mielessä puheenkoodaus on siis "*valmis*".

Edessä odottaa kuitenkin uusia jännittäviä ongelmia. Aalto-yliopistolla tutkitaan esimerkiksi nyt miten voitaisiin käyttää kaikkia lähellä olevia laitteita samanaikaisesti puheenkoodaukseen. Onhan selvää että jos kaikkia lähellä olevia mikrofoneja käyttää yhdessä, saadaan parempi äänenlaatu. Toisin kuin hands-free laitteissa, meidän tarkoitus on kuitenkin että mitä tahansa laitetta voisi käyttää puheenkoodaukseen, eikä sinun tarvitsisi erikseen ostaa ja pitää mukana hands-free kapinetta (kts. kuva 13).

Äänenlaadun lisäksi hajautettu puheenkoodaus parantaisi myös käyttöliittymää, koska puhetta ei tarvitsisi enää kohdistaa mihinkään tiettyyn laitteeseen vaan käyttäjä voisi vapaasti (vapaammin) liikkua tilassa. Suuri haaste on kuitenkin yksityisyys- ja turvallisuusasiat. Mikäli kaikki laitteet kaappaavat jatkuvasti ääntä, miten voimme taata käyttäjille yksityisyydensuojan? Tästä lisää luennolla.

4 Puheteknologian sovellus: Puheensiistaus/-ehostus

Puheteknologian laitteita käytetään oikeassa maailmassa, joka ei ole kliininen laboratorioympäristö, vaan ympärillämme on usein suuri määrä häiritseviä taustaääniä. Olet varmaan huomannut että liikenneruuhkassa tai yökerhossa on han-



Kuva 13: Hajautetun puheenkoodauksen perusidea; kaikki lähellä olevat laitteet osallistuvat puheenkoodauksen siten että äänenlaatu paranee.

kala puhua puhelimessa. Myös huonekaiku saattaa olla häiritsevä; huomaat sen kun puhut puhelimeen vessassa tai kun kuuntelet kuulutuksia rautatieasemalla. Häiriöt vaikeuttavat puheen ymmärtämistä, voivat joskus jopa estää ymmärretyksi tulemisen tai muuten vain rasittaa kuulijaa.

Ensimmäinen vaihtoehto häiriöiden helpottamisessa on low-tech lähestymistapa; mene pois melusta tai odota kunnes bussi ajaa ohi. Aina tämä ei ole mahdollista. Silloin olisi hyvä jos teknologia voisi auttaa.

Puheensiistauksella tai -ehostuksella viitataan tekniikoihin joilla poistetaan taustamelua tai huonekaikua, tai muutoin pyritään parantamaan puheäänen laatua tai ymmärrettävyyttä. Ehostuksen osa-alueita ovat mm. kohinanpoisto (jossa pyritään poistamaan muut äänet kuin haluttu ääni), lähde-erotus (jossa erotellaan äänet toisistaan), kaiunpoisto (jossa pyritään poistamaan huonekaiun vaikutus), sekä monimikrofonimenetelmät (jossa käytetään useampia mikrofoneja kohinan poistoon, lähde-erotukseen ja kaiunpoistoon).

Kohinanpoistomenetelmät (noise attenuation) perustuvat useimmiten oletamaan että kohina on additiivista, eli että puhesignaali s ja taustakohina v ovat toisistaan riippumattomia ja summautuvat havainnoksi

$$x = s + v. \tag{1}$$

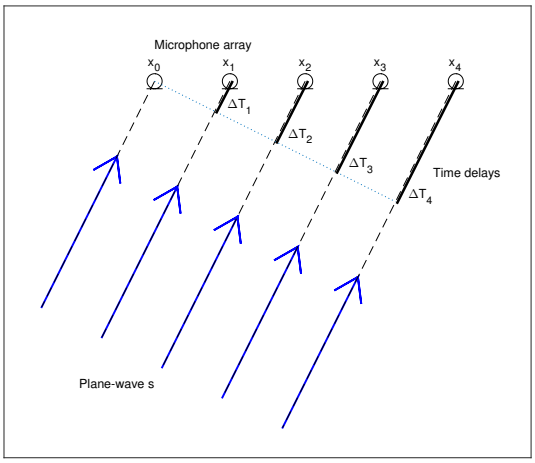
Tässä oletamme että x, s ja v ovat kompleksiarvoisia spektri- eli taajuuskomponentteja.

Jos edelleen oletamme että kohinasignaali ei juurikaan muutu ajassa (se on stationäärinen), niin voimme mitata sen keskimääräistä energiaa $|v|^2$ signaalista silloin kun puheäänessä on tauko. Siten voimme aproksimoida

$$\begin{aligned} |x|^2 &= |s+v|^2 \approx |s|^2 + |v|^2 \\ \Rightarrow |s| &\approx \sqrt{|x|^2 - |v|^2}. \end{aligned} \tag{2}$$

Kompleksiarvoisten muuttujien vaihetta (kompleksinen kulma) on kuitenkin vaikeampi mallintaa. Sen sijaan voimme olettaa että $|v|$ on pieni, joten

$$\angle x = \frac{x}{|x|} \approx \frac{s}{|s|} = \angle s. \tag{3}$$



Kuva 14: Keilanmuodostus monimikrofonimenetelmissä.

Tulokset yhdistämällä saamme estimaatin

$$\boxed{s \approx x\sqrt{\frac{|x|^2 - |v|^2}{|x|^2}}}. \tag{4}$$

Tämä tulos tarkoittaa sitä että mikäli kohina on huomattavasti hiljaisempi kuin puhesignaali, voimme laskea estimaatin puhesignaalille käyttäen ainoastaan arvausta kohinan energiasta. Mikä parempaa, estimaatin laskeminen on helppoa; havainto x pitää vain skaalata positiivisella kertoimella.

Kun käytössä on useampia mikrofoneja, voidaan niitäkin tietysti käyttää hyväksi signaalin ehostuksessa. Jos mikrofonit ovat eri etäisyydellä äänilähteestä, saapuu ääni eri mikrofoneihin eri aikaa. Sopivasti viivästämällä signaaleja, saadaan haluttu signaali kohdakkain jokaisella kanavalla siten että kun kanavat summaa toisiinsa, vahvistaa jokainen kanava haluttua signaalia. Muut äänilähteet eivät todennäköisesti ole täsmälleen kohdakkain, joten kun ne summaa toisiinsa, kumoavat ne toisensa ainakin osittain. Tämä metodi, joka tunnetaan nimellä delay-and-sum, on yksinkertainen keilanmuodostusmenetelmä (beamforming method) joka toimii sekä kohinan- että kaiunpoistoon. Sitä voi kuitenkin helposti parantaa siten että keilan suuntaavuus on parempi ja häiriöt vähenevät siten tehokkaammin.

Ehostusalgoritmeja käytetään nykyisin laajasti esimerkiksi kännyköissä. Usein kännyköissä on kaksi tai useampia mikrofoneja, joista etu- ja takapuolella on molemmilla ainakin yksi. Näin toinen on aina puhujan puolella ja toista voi käyttää kohinan estimointiin. Autoihin asennettavat hands-free laitteet käyttävät myös keilanmuodostusta, siten että esimerkiksi vain kuljettajan ääni välittyy puhelimelle ja sekä auton melu että muiden matkustajien äänet häivytetään.

(a) (b) (c)

Kuva 15: Puheteknologian professorit Aalto-yliopiston Signaalinkäsittelyn ja akustiikan laitoksella. (a) Professor of practice Tom Bäckström, (b) Akatemiaprofessori Paavo Alku, sekä (c) Professori Mikko Kurimo.

5 Puheteknologia Aalto-yliopistolla

Puheteknologian tutkimus ja opetus on Aalto-yliopistolla keskittynyt Signaalinkäsittelyn ja akustiikan laitokselle, jossa työskentelee tätä nykyä parisenkymmentä tutkijaa professorien Paavo Alku, Mikko Kurimo ja Tom Bäckström alaisuudessa. Laitoksen puheteknologian kurssit käsittävät mm:

ELEC-E5500	Speech Processing
ELEC-E5510	Speech Recognition L
ELEC-E5550	Statistical Natural Language Processing L
ELEC-E5520	Speech and Language Processing Methods L
ELEC-E5530	Speech and Language Processing Seminar L V

Viitteet

[1] J Benesty, M Sondhi, and Y Huang. *Springer Handbook of Speech Processing*. Springer, 2008.

[2] T Bäckström. *Speech Coding with Code-Excited Linear Prediction*. Springer, 2017.