Puheentunnistus

Mikko Kurimo

Teknillinen korkeakoulu
Informaatiotekniikan laboratorio

1 Johdanto

1.1 Puheentunnistuksen merkitys

Puhetta ymmärtävää konetta on pidetty tärkeänä askeleena ihmisen arkielämää helpottamaan kehitetyn teknologian kehityksessä. Sen avulla monet muut tekniikan saavutukset saadaan käyttöön ilman käsin tai muuten tapahtuvaa yksityiskohtaista ohjausta, ikäänkuin inhimillisen palvelijan avulla, joka viisaasti täyttää herransa toiveet. Monessa tapauksessa on kuitenkin osoittautunut kätevämmäksi ohjata koneiden toimintaa käsin, useimmiten niin että ihminen opettelee jonkin uuden taidon kuten autolla ajon tai tietokoneen ohjelmoinnin. Tästä huolimatta puheella ohjattavaa konetta pidetään teknologian saavutuksena aivan erityisessä arvossa. Usein tällainen kone kuvitellaan jollain tavalla tavallista älykkäämmäksi ja pystyvämmäksi suorittamaan vaativia tehtäviä joita käyttäjän ei tarvitse tai hän ei osaa yksityiskohtaisesti määritellä.

Suuri merkitys puheella ohjattavalla koneella on myös vammautuneiden tai muiden sellaisten ihmisten käytössä, joilla on vaikeuksia selvitä joistakin tavallisista nykyelämän tehtävistä. Juuri tämä ajatus ihmisen ja koneen välisen vuorovaikutuksen mahdollisesta helpottumisesta selittää automaattisen puheentunnistuksen saavuttamaa erityisasemaa teknologian kehittämisessä. Lisäksi asemaa korostaa se, ettei puheentunnistusongelmaan ole toistaiseksi saavutettu tyydyttävää ratkaisua, mittavista yrityksistä huolimatta.

1.2 Puheentunnistuksen vaikeudet

Mikä sitten tekee puheesta niin vaikeaa automaattisesti tunnistettavaksi ja herkkää erilaisille häiriöille ja puheen ja olosuhteiden muutoksille? Ensinnäkin puhesignaali on luonteeltaan jatkuvaa eikä ole itsestäänselvää, miten siitä eroitellaan yksittäiset sanat, lauseet ja puheenvuorot. Aina ei ole edes automaattisesti helppoa erottaa puhetta musiikista ja muista äänistä vaikkeivät ne olisikaan signaalissa päällekäin. Luonnollinen puhe sisältää myös runsaasti erilaisia muita tuotettuja ääniä, joita ei ole tarkoitus muuttaa tekstiksi, mutta niiden automaattinen erottaminen puheesta tuottaa vaikeuksia. Puheessa esiintyy lisäksi paljon vaihtelevuutta puhenopeuden ja ääntämisen suhteen eri puhujilla ja yhdelläkin puhujilla eri sanoissa ja konteksteissa. Kielessä ongelmia tuottavat sanat jotka kuulostavat lähes samoilta tai jopa täsmälleen samoilta ja tunnistimelle opetetun sanaston ulkopuoliset sanat, kuten erisnimet ja vieraskieliset sanat. Akustisissa olosuhteissa automaattiseen tunnistukseen vaikuttavat melun lisäksi tallennuslaitteiston, kuten mikrofonin ja välityskanavien ominaisuudet ja huoneen tai ulkotilan akustiikka. Yhteenvetona näistä kaikista ongelmista voi todeta että ihmiskorva ja aivojen kuuloalue ja informaation käsittely ovat ilmeisesti aina olleet ihmisen selviämiselle niin tärkeitä että niiden toiminta on kehittynyt erittäin robustiksi eri olosuhteisiin. Siksi myös automaattiselle tunnistuksellekin helposti asetetaan erittäin korkeita toimintavaatimuksia verrattuna muihin teknisiin apuvälineisiin.

2 Puheentunnistusongelman matemaattinen mallinnus

Automaattisen puheentunnistuksen lähestymistapa on tyypillinen hahmontunnistusongelma. Aiemmin tallennetuista signaalinäytteistä on oppivilla ja tilastollisilla menetelmillä estimoitu malleja, joihin uutta mitattua signaalia verrataan. Tunnistustulokseksi valitaan sitten teksti, joka vastaa sitä puhuttua viestiä jonka signaalia vastaavat mallit olisivat suurimmalla todennäköisyydellä tuottaneet. Matemaattisesti tehtävää kuvataan usein Bayesin kaavan avulla:

$$\Pr(W|X,M) = \frac{\Pr(X|W,M)\Pr(W|M)}{\Pr(X|M)}, \tag{1}$$

jossa W symboloi puheen avulla välitettyä viestiä, X mitattua signaalia ja M estimoitujen mallien joukkoa. Tunnistustulokseksi valitaan siis maksimitodennäköisyyttä vastaava teksti W^*.

Kaavassa 1 todennäköisyys $\Pr(X|W,M)$ lasketaan sovittamalla mittaussignaalia malleihin ja vaihtoehtoisiin viestihypoteeseihin. Todennäköisyys $\Pr(W|M)$ sensijaan sisältää viestihypoteesin apriori todennäköisyyden ns. kielimallin avulla, siis riippumatta mitatusta puhesignaalista. Mittaussignaalin kokonaistodennäköisyyden $\Pr(X|M)$ laskemista ei välttämättä tarvita haettaessa vain tulosta W^*, joka maksimoi todennäköisyyden $\Pr(W|X,M)$.

2.1 Ongelman hierarkkinen osittaminen

Puheentunnistuksen vaikeus tulee esiin rakennettaessa kaavaan (1) sopivia matemaattisia malleja (M). Äänenä havaittavien fysikaalisten paineaaltojen ja siinä välittyvän kielellisen viestin yhteyttä on toistaiseksi mahdotonta formuloida matemaattisesti, sillä kaikkia siinä olevia riippuvuuksia ja siihen vaikuttavia ilmiöitä ei tarkoin tunneta. Yleisesti käytössä oleva hierarkkinen ajattelutapa jakaa viestin sanoiksi, sanat edelleen foneemeiksi ja foneemit esimerkiksi prosesseiksi, jolla ihmisen ääntöväylä niitä tuottaa. Tarkempi tutkimus on kuitenkin osoittanut että nämä välivaiheet ovat vain ilmiöiden karkeita yksinkertaistuksia eikä tällä tavoin irrallisista osista rakennetuilla malleilla voida saavuttaa täydellistä tunnistustulosta.

Puheentunnistuksen tekee erityisen mielenkiintoiseksi se että on olemassa biologinen systeemi joka kuitenkin kykenee tunnistamaan puhetta erittäin hyvin monenlaisista häiriötekijöistä huolimatta. Automaattisten puheentunnistusmenetelmien kehityksessä tämä malli on otettu jo varhain huomioon ja pyritty etsimään laskentamalleja, jotka käyttäisivät hyväkseen joitakin samoja periaatteita mitä ihmisaivotkin noudattavat tietojenkäsittelyssään. Niinpä puheentunnistus on pitkään ollut erilaisten uusimpien älykkäiden laskentamenetelmien, kuten neuraalilaskennan algoritmien, testipenkkinä ja näitä onkin menestyksellä käytetty monessakin eri puheentunnistusprosessin osavaiheessa, joissa tavanomaiset matemaattiset laskenta- ja mallinnusmenetelmät ovat osoittautuneet tehottomiksi.

Vaikka on tunnettua ettei puheen hierarkkinen jako sanojen ja foneemien malleiksi tuottaisikaan tarkastiottaen parasta mahdollista lopputulosta, tämä tehtävän ja mallien pilkkominen osiinsa on kuitenkin hyväksytty lähtökohdaksi automaattiselle puheentunnistukselle. Syy on yksinkertaisesti se että tämänkaltaiselle rakenteelle on löydettävissä tehokkaita matemaattisia ratkaisumenetelmiä, jotka mallin likimääräisyydestä huolimatta voivat joissakin tapauksissa tuottaa tyydyttävän tunnistustuloksen.

2.2 N-gram -malli

Yksinkertainen matemaattinen malli puheviestin välityksessä käytetylle kielelle on ns. sana-n-gram, jossa jokaisella sanayhdistelmällä on tietty todennäköisyys. Tämän mallin avulla kunkin sanasekvenssissä esiintyvän sanan w_k todennäköisyys on laskettavissa riippuen $n-1$ edellisestä sanasta $w_{k-1}, \ldots, w_{k-n+1}$:

$$\Pr(w_k | w_{k-1}, w_{k-2}, \ldots, w_1) = \Pr(w_k | w_{k-1}, \ldots, w_{k-n+1}) \,. \tag{2}$$

Koko sanasekvenssin $W = w_1, w_2, \ldots, w_T$ todennäköisyydeksi muodostuu sitten, esimerkiksi puheentunnistuksessa tavallisen 3-grammin (trigrammin) avulla

$$\Pr(W) = \Pr(w_1) \Pr(w_2 | w_1) \prod_{k=3}^{T} \Pr(w_k | w_{k-1}, w_{k-2}) \,. \tag{3}$$

Luonnollisesti kaikille harvinaisille sana-n-grammeille ei voida, eikä ole järkevääkään, estimoida omia todennäköisyyksiään vaan niiden kohdalla sovelletaan $n-1$, ja tarvittaessa $n-2$ jne., todennäköisyyksiä. Käytännössä n-grammitodennäköisyydet on muutenkin järkevää tasoittaa vastaavien $1, 2, \ldots, n-1$ -grammien painotetulla summalla, jossa painot estimoidaan opetusaineiston kattavuuden perusteella [4]. Rajoittavina oletuksina tässä matemaattisessa mallissa on ettei n:ää sanaa kauemmas kantavia riippuvuuksia huomioida ja että sanasekvenssien todennäköisyydet ovat yksikäsitteisesti määrättävissä koko kielen sovellusalueella.

2.3 Gaussian mixture- ja kätketty Markov-malli

Sanojen ääntymisen yksinkertaisin malli on kuvata sanat foneemijonoina, joissa sanan akustinen todennäköisyys tietylle signaalille saadaan suoraan foneemijonon todennäköisyydestä. Foneemit esiintyvät äänteinä, joiden akustisille havainnoille käytetty matemaattinen malli on ns. kätketty Markov-malli (HMM). Siinä äänteitä tuottavan systeemin oletetaan koostuvan tiloista, joissa syntyy tilastollisilta ominaisuuksiltaan tiettyä stationääristä todennäköisyysjakaumaa noudattavaa signaalia. Todennäköisyys, jolla tila i tuottaa signaalia kuvaavan piirrevektorin $\boldsymbol{x}$, saadaan esimerkiksi ns. Gaussian mixture -mallista (GMM):

$$b_i(\boldsymbol{x}) = \sum_{j=1}^{J} c_{ij} b_{ij}(\boldsymbol{x}) \,, \tag{4}$$

jossa mikstuurin painot täyttävät ehdot: $c_{ij} \geq 0$ ja $\sum_{j=1}^{J} c_{ij} = 1$. Mikstuurikomponentit ovat tavallisesti moniulotteisia normaalijakaumia $b_{ij}(\boldsymbol{x}) \sim N(\mu_{ij}, \Sigma_{ij})$.

Systeemin tilojen vaihtumista säätelee toinen stokastinen prosessi, jossa kaikille mahdollisille tilasiirtymille on estimoitu oma siirtymätodennäköisyytensä. Tämä jälkimmäinen prosessi on "kätketty" ulkopuolisilta havaitsijoilta, sillä tilasiirtymiä ei voi suoraan havaita vaan ne näkyvät ulospäin ainoastaan tuotetun signaalin tilastollisten ominaisuuksien muutoksina. Karkeastiottaen tällaista tilaa voidaan verrata esimerkiksi ihmisen ääntöväylän tiettyyn asentoon, joka ei ole ulospäin näkyvissä, mutta joka kuullaan tietynlaisena tuotettuna äänenä. Näistä peräkkäisistä äänteistä koostuvat sitten foneemit ja peräkkäisistä foneemeista sanoja vastaavat foneemijonot. Yhteistodennäköisyys, jolla malli M tuottaa tilajonon $Q = q_0, \ldots, q_T$ ja sillä havaintosekvenssin $\boldsymbol{X} = \boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$, on laskettavissa kaavasta

$$\Pr(\boldsymbol{X}, q|M) = \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(\boldsymbol{x}_t) \,. \tag{5}$$

Koska havainnot generoiva tilajono on tuntematon, havaintosekvenssin todennäköisyys mallille M on summa kaikkien mahdollisten tilajonojen yli

$$\Pr(\boldsymbol{X}|M) = \sum_{q} \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(\boldsymbol{x}_t) \,. \tag{6}$$

Käytännössä (6) lasketaan ns. forward-backward -menetelmällä [3]. Rajoittavina oletuksina tässä matemaattisessa mallissa on että systeemi voi olla vain yhdessä tilassa kerrallaan ja että siirtymistodennäköisyys seuraavaan tilaan riippuu vain edellisestä tilasta. Lisäksi sanojen koostuminen foneemijonoista ja foneemien tilajonoista on oltava yksikäsitteisesti määritettävissä.

3 Hahmontunnistusongelma ja sen ratkaisu

3.1 Mallin rakenne

Tyypillinen state-of-art puheentunnistusjärjestelmä koostuu seuraavista yhteennivoutuvista malleista:

1. **Kielimalli** antaa todennäköisyydet sanoille ja sanayhdistelmille perustuen yleensä suureen kieliaineistoon, jossa tehtävään liittyvä sanasto ja sanontatavat esiintyvät oikeissa tilastollisissa suhteissa. Joissakin suppean sanaston erityissovelluksia nämä todennäköisyydet voidaan olettaa myös tehtävän yhteydessä määrätyiksi. Tavallinen laajan sanaston tunnistuksessa käytetty malli on edellämainittu n-gram (kaava 3), jossa periaatteessa kullekin n:n sanan yhdistelmälle on estimoitu oma esiintymistodennäköisyytensä. Käytännössä riittävän opetusaineiston järjestäminen mallin parametrien estimoimiseen on yleensä mahdotonta, joten älykkäiden oppimismenetelmien käyttö hyvien mallien saamiseksi on välttämätöntä.

2. **Sanastomalli** (leksikko) kertoo mitä sanoja on olemassa ja mistä foneemeista ne koostuvat. Joillakin sanoilla voi myös olla useita mahdollisia ääntämistapoja, jolloin niille on estimoitava esiintymistodennäköisyydet. Suomenkielessä sanat voidaan muuttaa foneemijonoiksi melko yksikäsitteisten sääntöjen avulla, mutta esimerkiksi englannissa, ja suomen vierasperäisten sanojen kohdalla, ääntämistavat on yleensä määritettävä käsin.

3. **Foneemimallin** avulla voidaan laskea todennäköisyydet, joilla puheesta erotettu signaalisegmentti olisi peräisin tietystä foneemista. Yhden foneemin malli koostuu yleensä peräkkäisistä tiloista, joissa tuotetun signaalin tietyt ominaisuudet oletetaan stationäärisiksi ja näille ns. akustisille piirteille voidaan estimoida tiheysfunktiomallit. Tyypillinen tällainen malli, jossa on erikseen mallinnettu piirteiden tiheysfunktiot systeemin eri tiloissa ja tilojen välisten siirtymien todennäköisyydet on edellä mainittu HMM (kaava 6).

3.2 Tyypillisen tunnistusprosessin vaiheet

Puheentunnistus automaattisen järjestelmän avulla etenee yleensä seuraavasti:

1. **Esikäsittely**. Mikrofonilla tallennettu signaali digitoidaan, jaetaan lyhyisiin noin kymmenen millisekunnin pituisiin ikkunoihin ja kullekin signaali-ikkunalle erikseen tehdään spektrianalyysi taajuustason informaation havaitsemiseksi. Piirteiden laskentaan on erilaisia hyväksihavaittuja algoritmeja, joille on yhteistä tiettyjen taajuuskaistojen tehojen mittaus ja tehopiikkien jaksollisuus. Tavallisin menetelmä on laskea tehospektristä diskreetti kosinimuunnos ja käyttää ihmiskorvan taajuusherkkyydestä johdettua MEL-asteikkoa tunnistuksen kannalta mielenkiintoisimpien kaistojen valintaan.

2. **Havaintojen kuvaaminen piirrevektorilla**. Foneemimallien syötteeksi valitaan kutakin aikajaksoa kohti piirrevektori ($\boldsymbol{x}$, kaavassa 4), joka spektrianalyysistä johdettujen kertoimien lisäksi sisältää jonkin verran tietoa naapuri-ikkunoista, esimerkiksi kertoimien muutosta kuvaavia tunnuslukuja. Piirrevektorin dimensio on tyypillisesti 20 – 100.

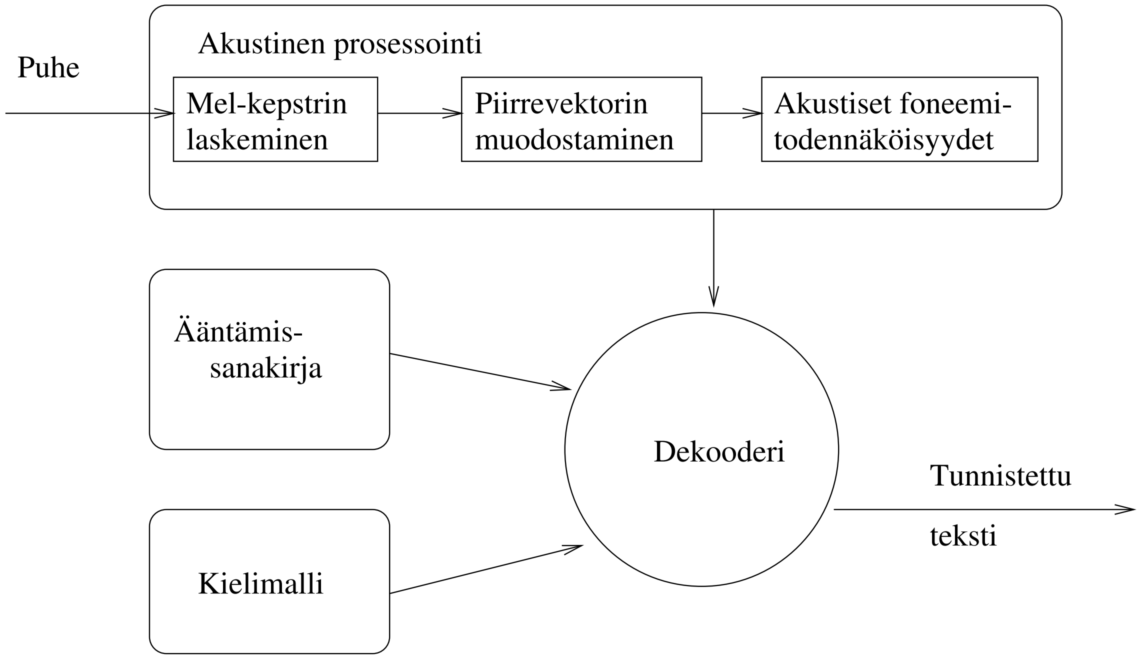
3. **Todennäköisyyksien laskeminen**. Vertaamalla foneemimalleihin liittyviä piirteiden tiheysfunktiomalleja havaittuihin piirrevektorisekvensseihin lasketaan kullekin foneemille (itseasiassa sen tiloille i) todennäköisyyssekvenssi, joka kuvaa todennäköisyyttä ($b_i(\boldsymbol{x})$, kaavassa 4) kullakin ajanhetkellä.

4. **Puheen dekoodaus**. Haetaan todennäköisin puheen sisältöä vastaava sanasekvenssi yhdistämällä dekooderin avulla havaitut foneemitodennäköisyydet sekä kielimallin antamat sanojen ja sanasekvenssien todennäköisyydet. Tuloksen käyttötarkoituksesta riippuen dekooderilla voidaan yhden tuloksen lisäksi laskea myös järjestetty listan muista todennäköisistä tulovsaihtoehdoista.

4 Nykyisten puheentunnistimien suorituskyky

4.1 Evaluointitehtävien luokittelu

Puheentunnistustehtävien laajasta vaihtelusta johtuen toiminnan tarkkuutta vertailtaessa on tärkeää korostaa tehtävän laatua ja sen rajoituksia. Tehtävät jaetaan usein puhujien määrän, puhetavan eli sanojen välisen tauotuksen ja puheen luonnollisuuden suhteen erilaisiin luokkiin. Käytettävissä olevaan kielen vaikeuden luokitusperusteita ovat sanaston koko ja kielen syntaksin rajoittavuus. Myös akustisten olosuhteiden perusteella tehtävät jaoitellan luokkiin kohinan, häiriöiden ja olosuhteiden stabiilisuuden mukaan.



Kuva 1: Puheentunnistusjärjestelmän toimintakaavio.

4.2 Laajan sanaston tunnistimien suorituskyvystä

Parhaiden nykyisten (englanninkielisten) puheentunnistimien tunnistustarkkuudeksi on mitattu esimerkiksi tavallisille radion ja television uutislähetyksille, keskimäärin 20 % sanavirheitä [2]. Sanavirheet tarkoittavat koko puhelähetyksen tunnistustuloksen sovitusta todelliseen tekstiin niin että virheiksi lasketaan hävinneet, ylimääräiset ja vaihtuneet sanat. Virheprosentti on virheiden määrä suhteessa todellisen tekstin sanamäärään. Tunnistuloksessa on tyypillistä että monet puheosuudet ovat lähes täysin virheettömiä, mutta joissakin hankalissa kohdissa tunnistusvirheiden määrä on suuri. Tunnistustulos on käyttökelpoinen esimerkiksi puhelähetysten indeksoinnissa, jonka avulla lähetyksistä voidaan etsiä sisällön perusteella kiinnostavia osia tai käyttää ääniaineistoa tiedonhaussa tekstiaineiston tapaan. Tämäntyyppisiä sovelluksia on käytössä muunmuassa tiedonhaussa uutisaineistoista, audio- ja videotallenteiden annotoinnissa ja suurien audioaineistojen indeksoinnissa. Täysin vapaaseen saneluun virhemäärä sen sijaan on usein liian suuri, mutta kun järjestelmän annetaan adaptoitua tietylle puhujalle ja tietyntyyppisille teksteille, tulos on verrattavissa tottumattomaan konekirjoittajaan, esimerkiksi 20 – 40 sanaa minuutissa sisältäen virheiden korjaukset [1].

4.3 Suppean sanaston tunnistimien suorituskyvystä

Rajoitetuissa erikoistehtävissä puheentunnistimien tarkkuus on yleensä huomattavasti parempi. Tehtävän rajoittaessa tilannekohtaisten sanavaihtoehtojen joukon pieneksi vaihtoehtojen akustiset erot ovat usein selkeitä ja tunnistin toimii lähes virheettömästi. Tunnistin voi suoriutua tehtävästään riittävän hyvin jopa lievästi häiriöalttiissa olosuhteissa ja suurelle joukolle eri puhujia. Näitä ominaisuuksia on menestyksellisesti hyödynnetty useissa (etupäässä englanninkielisissä) sovelluksissa, joissa yksinkertaisia puhelinpalveluita on automatisoitu, esimerkiksi aikatauluneuvontaa, paikallisia säätiedotuksia, urheilutulosten välitystä ja joitakin pankkipalveluita. Avainasemassa on älykäs käyttöliittymä, joka sopivalla tavalla ohjaa käyttäjää saamaan haluamansa tiedot niin että puheentunnistustehtävät jäävät helpoiksi palvelun ollessa silti riittävän joustavaa ja tehokasta. Muita vastaavia tehtäviä jotka joissakin tapauksissa sopivat hyvin puheen avulla ohjattaviksi ovat mm. vammaisten apuvälineet, handsfree-puhelimet ja erilaiset pienet vaatteissa kannettavat laitteet kuten audio- (ja video-) tallentimet ja soittimet ja mittalaitteet. Ongelmia tunnistuksessa aiheuttavat meluisat käyttötilanteet ja epäselvästi tai poikkeavasti puhuvat käyttäjät. Manuaalinen käyttö sopii puheentunnistusta paremmin myös kriittisiin sovelluksiin joissa edellytetään 100 %:sta toimintavarmuutta kaikissa tilanteissa.

5 Alan tutkimus Suomessa

5.1 Poimintoja tutkimuksen historiasta ja nykytilanteesta

Informaatiotekniikan laboratoriossa 1970-luvun lopussa rakennettu muutaman tuhannen sanan käsittävä automaattinen puheentunnistin lienee ensimmäinen suomenkielinen puheentunnistusjärjestelmä. Tunnistin perustui oppivaan aliavaruusmenetelmään ja redundanttiin hash-osoitukseen. Tämän jälkeen puheentunnistus on ollut laboratorion eräs tärkeimmistä uusien hahmontunnistusalgoritmien testipenkeistä, joilla algoritmien soveltuvuutta hankalan mittausdatan analyysissä on voitu mitata ja verrata state-of-art -menetelmiin. Samalla saavutetut hyvät tulokset ovat jopa maailmanlaajuisesti vaikuttaneet sekä puheentunnistuksen että itse algoritmien kehityssuuntiin. Tunnetuimpia puheentunnistukseen vaikuttaneita laboratoriossa akateemikko Teuvo Kohosen johdolla kehitettyjä algoritmeja ovat olleet itseorganisoiva kartta (1981) [5] ja oppiva vektorikvantisaatio (1986) [6]. Puheentunnistuksen merkkipaaluja laboratoriossa puolestaan ovat olleet mm. foneettinen kirjoituskone (1988) [7] ja diskreetteihin kätkettyihin Markov malleihin (HMM) perustuva rajattoman sanaston tunnistin (1991) [9].

Nykyään hahmontunnistustutkimusta Suomessa tehdään puheentunnistuksen piirissä monella tasolla. Opetusministeriö teetti vastikään (2001) [8] esiselvityksen puheentutkimuksen resursseista Suomessa, jolla pyrittiin selvittämään puheentutkimuksen nykytilaa ja toimia tutkimusedellytysten parantamiseksi. Puheentutkimukseen liittyy paljon muitakin tutkimuskohteita kuin puheentunnistus, mutta puheentunnistus on kuitenkin tärkeällä sijalla, sekä itsenäisenä tutkimusalana että monen muun alan tarvitsemana työkaluna. Lisäksi puheentunnistus merkittävänä hahmontunnistusongelmana on noussut maailmanlaajuisesti tärkeäksi hahmontunnistusalgoritmien benchmark-testausvälineeksi, jolla monien uusia algoritmien suorituskykyä voidaan evaluoida state-of-art -menetelmiin.

5.2 Puheen signaalinkäsittely

Piirteiden irroitus puhesignaalista on kaikissa state-of-art puheentunnistusjärjestelmissä melko samanlainen. Kuitenkin kun tällaista lyhyestä aikaikkunasta laskettuihin spektrogrammeihin pohjautuvaa menetelmää verrataan esimerkiksi ihmisen puheentunnistusmekanismin kykyyn löytää invariantteja piirteitä eri ihmisten puheesta erilaisissa olosuhteissa, on selvää että parantamisen varaa on huomattavasti. Uudenlaisia piirreirroitusmenetelmiä puheentunnistukseen tutkitaan tällä hetkellä aktiivisesti mm. *TKK:n akustiikan laboratoriossa.*

5.3 Kohinansieto, monikielisyys

Automaattisten menetelmien soveltuvuus käytännön puhetilanteisiin, joissa esiintyy monenlaisia akustisia häiriöitä ja kohinaa on viimeaikoina herättänyt yhä enemmän huomiota puheentunnistustutkimuksessa. Erityisesti mm. *Nokian tutkimuskeskuksessa* kehitetään robustia puheteknologiaa, jolla puhutut komennot voidaan tunnistaa ja ymmärtää oikein. Tämäntyyppiset puheentunnistustehtävät ovat suhteellisen kieliriippumattomia, jopa niin että voidaan kehittää myös monikielisiä tunnistusjärjestelmiä. *TTKK:n digitaalisen median instituutissa* tutkitaan mm. akustisia malleja, jotka sopivat puhelinkaistaan ja monikieliseen puheentunnistukseen.

5.4 Puhekäyttöliittymät

Puhtaasti puheen avulla toimiva käyttöliittymä on eräs puheteknologian mielenkiintoisimmista haasteista. Vuorovaikutus puheen avulla poikkeaa kuitenkin niin paljon perinteisistä kojetauluista, säätimistä ja näppäimistöistä, että järjestelmien käytettävyystutkimus muodostaa aivan oman tieteenalansa. Puhekäyttöliittymiä tutkitaan erityisesti *Tampereen yliopiston TAUCHI-tutkimusryhmässä.*

5.5 Kielimallit

Puheentunnistuksen laajempi käyttö tiedonvälityksessä edellyttää yksittäisiä sanoja tai komentolauseita laajempien kokonaisuuksien tunnistamista. Näissä tunnistustehtävissä korostuu kielen mallinnus eli morfologisten, syntaktisten ja semanttisten riippuvuuksien huomionti, koska sanat usein esiintyvät eri muodoissaan ja ovat riippuvaisia kontekstista. Lisäksi osia puheesta joudutaan usein arvailemaan sillä epäselvästi lausutut kohdat ovat automaattiselle tunnistimelle vaikeimpia. Kielimallit jäljittelevät osittain inhimillistä tapaa ratkaista tämä ongelma eli kontekstin mallin perusteella voidaan esittää todennäköisempiä hypoteesejä puheen sisällöstä, joita sitten verrataan mitattuun puhesignaaliin. *TKK:n neuroverkkojen tutkimusyksikössä* laskennallisesti tehokkaiden ja suuria opetusaineistoja hyödyntävien, adaptiivisten kielimallien kehitys liittyy läheisesti laajan sanaston jatkuvan puheentunnistuksen tutkimukseen. Tutkimusaiheisiin kuuluu myös näiden kielimallien avulla tapahtuva puheenaiheen karakterisointi ja tämän tiedon käyttö puheeseen sisältyvän viestin ymmärtämiseen puhedialogeissa.

5.6 Puheen mallien generointi ja adaptointi

Foneemien ja foneemisekvenssien tilastollisena mallina nykyään laajalti käytössä oleva HMM-malli on monessakin mielessä varsin epäsopiva näin vaativaan tehtävään. Malli on kuitenkin matemaattisesti erittäin kätevä ja sopivien laajennusten avulla se on saatu toimimaan kohtuullisen hyvin ja riittävän tehokkaasti monessa puheentunnistussovelluksessa. Siksi HMM:n laajennukset, tehostukset ja nopea adaptaatio, mm. neuraalilaskennan sovelluksena, on tutkimuskohteena *TKK:n neuroverkkojen* *tutkimusyksikössä*. Nykyään tutkimuskohteena on myös yleisemmät dynaamiset tilamallit aikasarjoille, joiden avulla HMM:n rajoituksista voidaan päästä eroon.

5.7 Puheen havaitseminen

Puhe on myös visuaalinen ilmiö ja ihmisen kyky integroida kuultu ja nähty puhe liittyy puheentunnistustutkimukseen. Visuaalinen puhe korostuu erityisesti tilanteissa, joissa esiintyy voimakkaita akustisia häiriöitä tai kuuloaistimus on muuten viallinen. *TKK:n laskennallisen tekniikan laboratoriossa* tutkitaan visuaalisen puheen havaitsemista ja käsittelyä. Kognitiivisia ihmiseen liittyviä tekijöitä auditiivisen puheteknologian puolella tutkitaan mm. *TKK:n akustiikan laboratoriossa*.

5.8 Kaupalliset puheentunnistustuotteet

Puheentunnistusteknologiaa on jo Suomessakin tuotu kaupallisten sovellusten avulla osaksi tavallisten ihmisten arkipäivää. Tästä on esimerkkejä mm. *Nokian* matkapuhelimissa, *Philipsin* puheentunnistusohjelmassa PC:lle ja *Soneran* ja *Elisan* kehittämissä automaattisissa puhelinpalveluissa. Myös pienemmissä yrityksissä on panostettu voimakkaasti suomenkielisten puheentunnistustuotteiden kehittämiseen (mm. *Lingsoft*).

5.9 Puheen indeksointi ja haku

Laajojen puheaineistojen tunnistustuloksia voidaan myös menestyksellisesti käyttää aineistojen indeksointiin ja siihen perustuvaan tiedonhakuun. Tästä on hyvinä esimerkkeinä amerikkalaiset SpeechBot- ja SpeechFind -järjestelmät, joilla internetin välityksellä voidaan hakea kiinnostava puheäänite valtavista arkistoista, jotka sisältävät jopa kymmeniä tuhansia tunteja materiaalia, esimerkiksi useampien vuosien radio-ohjelmat tai nauhoitetut julkiset puheet sadan vuoden ajalta. Tiedonhaku puheentunnistustulosten perusteella on käyttökelpoinen tapa myös audiovisuaalisen signaalin (esimerkiksi videon) käsittelyssä. Suomessa puheentunnistuksen avulla tapahtuva suurten aineistojen indeksointi on kiinnostuksen kohteena mm. *Oulun yliopiston MediaTeamissa* ja *TKK:n neuroverkkojen tutkimusyksikössä*.

Viitteet

[1] *PC Magazine*. December 1999.

[2] *Proceedings of DARPA Broadcast News Workshop*. NIST, 1999.

[3] L.E. Baum. An inequality and associated maximization technique in statistical estimation of probabilistic functions of Markov processes. *Inequalities*, 3:1–8, 1972.

[4] F. Jelinek and R.L. Mercer. Interpolated estimation of Markov source parameters from sparse data. In *Proceedings of an International Workshop on Pattern Recognition in Practice*, Amsterdam, The Netherlands, 1980. North-Holland.

[5] Teuvo Kohonen. Automatic formation of topological maps of patterns in a self-organizing system. In Erkki Oja and Olli Simula, editors, *Proc. 2SCIA, Scand. Conf. on Image Analysis*, pages 214–220, Helsinki, Finland, 1981. Suomen Hahmontunnistustutkimuksen Seura r.y.

[6] Teuvo Kohonen. Learning vector quantization for pattern recognition. Report TKK-F-A601, Helsinki University of Technology, Espoo, Finland, 1986.

[7] Teuvo Kohonen. The 'neural' phonetic typewriter. *Computer*, 21(3):11–22, 1988.

[8] Juhani Toivanen and Manne Miettinen. *Puheentutkimuksen resurssit Suomessa.* CSC - Tieteellinen laskenta Oy, Finland, 2001. (in Finnish).

[9] Kari Torkkola, Jari Kangas, Pekka Utela, Sami Kaski, Mikko Kokkonen, Mikko Kurimo, and Teuvo Kohonen. Status report of the Finnish phonetic typewriter project. In T. Kohonen, K. Mäkisara, O. Simula, and J. Kangas, editors, *Artificial Neural Networks*, volume I, pages 771–776, Amsterdam, Netherlands, 1991. North-Holland.