PUHEENTUNNISTUS

Mikko Kurimo, Teknillinen korkeakoulu,
Tietojenkäsittelytieteen laitos,
Adaptiivisen Informatiikan tutkimusyksikkö

Automaattinen puheentunnistus on merkittävä tilastollisten ja oppivien hahmontunnistusmenetelmien sovellus, joka on erityisasemassa myös lukuisten tavallisillekin ihmisille läheisten sovellustensa vuoksi. Vaikka puheentunnistus on ihmiselle helppoa, on se kuitenkin erittäin haastava ongelma äänisignaalin ja kielen rikkauden ja monimuotoisuuden vuoksi. Tässä artikkelissa esitellään lyhyesti nykyaikaisen puheentunnistimien toimintaperiaatetta, ratkaisujen matemaattista perustaa sekä tunnistinten suorituskykyä. Lisäksi luodaan lyhyt katsaus puheentunnistuksen tutkimukseen Suomessa.

Avainsanat: Puheentunnistus.

JOHDANTO

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, ikään kuin 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.

Kirjoittajan yhteystiedot:
Adaptiivisen Informatiikan tutkimusyksikkö,
Tietojenkäsittelytieteen laitos,
Teknillinen korkeakoulu/
Adaptive Informatics Research Centre,
Department of Information and Computer Science
Helsinki University of Technology
PL 5400, 02015 TKK
Sähköposti Mikko.Kurimo@tkk.fi

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 itsestään selvää, miten siitä erotellaan yksittäiset sanat, lauseet ja puheenvuorot. Aina ei ole automaattisesti helppoa erottaa puhetta edes musiikista ja muista äänistä, vaikkeivät ne olisikaan signaalissa päällekkäin. Luonnollinen puhe sisältää myös runsaasti erilaisia muita tuotettuja ääniä kuten epäröintejä, korjauksia, yskintää, joita ei tunnistuksessa ole tarvetta muuttaa tekstiksi, mutta niiden automaattinen erottaminen puheesta tuottaa vaikeuksia. Eri puhujien puheessa esiintyy lisäksi paljon vaihtelevuutta puhenopeuden ja ääntämisen suhteen ja samallakin puhujalla usein eri sanoissa ja konteksteissa. Kielessä ongelmia tuottavat sellaiset sanat, jotka kuulostavat lähes samoilta tai jopa täsmälleen samoilta, sekä tunnistimelle opetetun sanaston ulkopuoliset sanat, kuten erisnimet, vieraskieliset sanat ja sanojen harvinaiset taivutusmuodot. 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 tunnistukselle asetetaan helposti erittäin korkeita toimintavaatimuksia verrattuna muihin teknisiin apuvälineisiin.

PUHEENTUNNISTUSONGELMAN MATEMAATTINEN MALLINNUS

Automaattisessa puheentunnistuksessa nykyisin käytetty lähestymistapa on tyypillinen tilastolliselle hahmontunnistusongelmalle. Aiemmin tallennetuista signaalinäytteistä estimoidaan oppivilla ja tilastollisilla menetelmillä 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 (1) avulla:

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

Kaavassa 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 $\Pr(W|M)$ sen sijaan sisältää viestihypoteesin apriori todennäköisyyden ns. kielimallin avulla - siis riippumatta mitatusta puhesignaalista. Kielimalli on opetusaineistosta tilastollisesti tai sovelluksen perusteella määrätty lauseiden yleinen todennäköisyysjakauma. Mittaussignaalin kokonaistodennäköisyyden $\Pr(X|M)$ laskemista, eli summaa yli kaikkien mahdollisten lauseiden, ei välttämättä tarvita haettaessa vain lausetta W^*, joka maksimoi todennäköisyyden $\Pr(W|X,M)$, eli on todennäköisin tunnistustulos.

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, joilla 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 kuitenkin olemassa biologinen systeemi, joka 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 monissa eri puheentunnistusprosessin osavaiheissa, joissa tavanomaiset matemaattiset laskenta- ja mallinnusmenetelmät ovat osoittautuneet tehottomiksi.

Vaikka on tunnettua, ettei puheen hierarkkinen jako sanojen ja foneemien malleiksi tuotakaan tarkasti ottaen 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.

N-gram-malli

Yksinkertainen matemaattinen malli puheviestin välityksessä käytetylle kielelle on niin sanottu sana-*n*-gram, jossa jokaisella sanayhdistelmällä on tietty todennäköisyys. Tämän mallin avulla kunkin sanasekvenssissä esiintyvän sanan *wk* todennäköisyys on laskettavissa riippuen $n-1$ edellisestä sanasta $wk-1, \ldots, wk-n+1$:

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

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

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

Luonnollisesti kaikille harvinaisille sana-*n*-grammeille ei voida, eikä ole järkevääkään, estimoida omia todennäköisyyksiään, vaan niiden kohdalla sovelletaan todennäköisyyksiä $n-1$, ja tarvittaessa $n-2$ jne. Käytännössä n-grammitodennäköisyydet on muutenkin järkevää tasoittaa vastaavien 1, 2,..., $n-1$ -grammien painotetulla summalla, jossa painot estimoidaan mallin opetusaineiston kattavuuden perusteella (Jelinek & Mercer, 1980). Rajoittavana oletuksena tässä matemaattisessa mallissa on, ettei *n*:ää sanaa kauemmas kantavia riippuvuuksia huomioida ja että sanasekvenssien todennäköisyydet määräytyvät kielessä yksikäsitteisesti eli eivät riipu muun muassa puheenaiheesta ja puhetilanteesta.

Gaussian mixture- ja kätketty Markov-malli

Yksinkertaisin tapa mallintaa sanojen ääntymistä, 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 (*Hidden Markov Model*, HMM). Siinä äänteitä tuottavan systeemin oletetaan koostuvan tiloista, joissa syntyy tilastollisilta ominaisuuksiltaan tiettyä stationääristä todennäköisyysjakaumaa noudattavaa signaalia.

Siis kunkin foneemin malli koostuu peräkkäisistä tiloista ja yhden tilan tuottama äänisignaali on piirteiltään tietynlaista. Todennäköisyys, jolla tila i tuottaa signaalia kuvaavan piirrevektorin x, saadaan esimerkiksi ns. Gaussian mixture -mallista (GMM):

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

jossa mikstuurin painot täyttävät ehdot:

$$c_{ij} \geq 0 \text{ ja } \sum_{j=1}^{J} c_{ij} = 1 .$$

Mikstuurikomponentit ovat tavallisesti moniulotteisia normaalijakaumia $b_{ij}(x) \sim N(\mu_{ij}, \Sigma_{ij})$.

HMM-systeemin tilojen vaihtumista säätelee toinen stokastinen prosessi, jossa kaikille mahdollisille tilasiirtymille on estimoitu oma siirtymätodennäköisyytensä. Yksittäisistä siirtymistä voidaan näin laskea kokonaisten tilaketjujen todennäköisyydet ja vertailla niitä toisiinsa. Tilojen vaihtumista säätelevä prosessi on kuitenkin "kätketty" ulkopuolisilta havaitsijoilta, sillä tilasiirtymiä ei voi suoraan havaita, vaan ne näkyvät ulospäin ainoastaan tuotetun signaalin tilastollisten ominaisuuksien muutoksina. Karkeasti ottaen 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 $X = x_1, \ldots, x_T$, on laskettavissa kaavasta

$$\Pr(X, q|M) = \pi_{q0} \prod_{t=1}^{T} a_{q\,t-1\,q\,t} b_{q\,t}(x_t) . \quad (5)$$

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

$$\Pr(X|M) = \sum_{q} \pi_{q0} \prod_{t=1}^{T} a_{q\,t-1\,q\,t} b_{q\,t}(x_t) . \quad (6)$$

Käytännössä (6) lasketaan ns. forward-backward –menetelmällä (Baum, 1972), jossa dynaamisen ohjelmoinnin avulla tehokkaasti haarukoidaan kunkin tilajonon todennäköisyys samanaikaisesti. Rajoittavina oletuksina tässä matemaattisessa mallissa ovat, 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 voitava määritellä yksikäsitteisesti.

HAHMONTUNNISTUSONGELMA JA SEN RATKAISU

Mallin rakenne

Tyypillinen nykyaikainen puheentunnistusjärjestelmä koostuu seuraavista yhteen nivoutuvista malleista:

1. *Kielimalli* antaa todennäköisyydet sanoille ja sanayhdistelmille. Kielimalli perustuu yleensä suureen kieliaineistoon, jossa tehtävään liittyvä sanasto ja sanontatavat esiintyvät oikeissa tilastollisissa suhteissa. Joissakin suppean sanaston erityissovelluksissa nämä todennäköisyydet voidaan olettaa myös tehtävän yhteydessä käsin määritetyiksi. Laajan sanaston tunnistuksessa tavallisesti käytetty malli on edellä mainittu n-gram (kaava 3), jossa periaat-

teessa kullekin *n*:n sanan yhdistelmälle on estimoitu oma esiintymistodennäköisyytensä. Riittävän opetusaineiston järjestäminen mallin parametriarvojen tarkkaan estimoimiseen on yleensä käytännössä mahdotonta, joten älykkäiden oppimismenetelmien käyttö hyvien ja yleistämiskykyisten mallien tuottamiseksi 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. Suomen kielessä sanat voidaan muuttaa foneemijonoiksi melko yksikäsitteisten sääntöjen avulla, mutta esimerkiksi englannissa ja suomen vierasperäisten sanojen kohdalla ääntämistavat on yleensä määritettävä käsin. Kielissä joissa sanat taipuvat voimakkaasti ja esiintyvät yleisesti erilaisten päätteiden ja etuliitteiden kanssa ja yhdyssanoina, käytetään sanojen sijasta usein lyhyempiä kielen yksikköjä kuten morfeemeja. Kielimalli voidaan opettaa morfeemeille samaan tapaan kuin sanoillekin.

3. *Foneemimallin* avulla voidaan laskea todennäköisyydet, joilla puheesta erotetut signaalisegmentit ovat peräisin tietyistä foneemeista. Yhden foneemin malli koostuu yleensä peräkkäisistä tiloista, joissa tuotetun signaalin spektristä lasketut piirteet oletetaan stationäärisiksi, ja näille (akustisille) piirteille voidaan estimoida tiheysfunktiomallit. Tyypillinen tällainen malli, jossa sekä piirteiden tiheysfunktiot systeemin eri tiloissa että tilojen välisten siirtymien todennäköisyydet on mallinnettu erikseen, on edellä mainittu HMM (kaava 6). Koska foneemit usein kuulostavat erilaisilta eri sanayhteyksissä, foneemimallit määritellään yleensä trifonimalleina, jolloin mallin parametrit riippuvat lisäksi sanassa edeltävästä ja seuraavasta foneemista.

Tyypillisen tunnistusprosessin vaiheet

Puheentunnistus automaattisen järjestelmän avulla etenee yleensä seuraavien vaiheiden kautta:

1. *Esikäsittely*. Mikrofonilla tallennettu signaali digitoidaan ja jaetaan lyhyisiin, noin kymmenen millisekunnin pituisiin ikkunoihin, minkä jälkeen kullekin signaali-ikkunalle tehdään erikseen spektrianalyysi taajuustason informaation havaitsemiseksi. Piirteiden laskentaan on erilaisia hyväksi havaittuja 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 (x, kaavassa 4), joka spektrianalyysistä johdettujen kertoimien lisäksi sisältää jonkin verran tietoa naapuri-ikkunoista, esimerkiksi kertoimien muutosta kuvaavia tunnuslukuja. Näistä kertoimista ja tunnusluvuista kootun piirrevektorin dimensio on tyypillisesti 20 – 100.

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

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

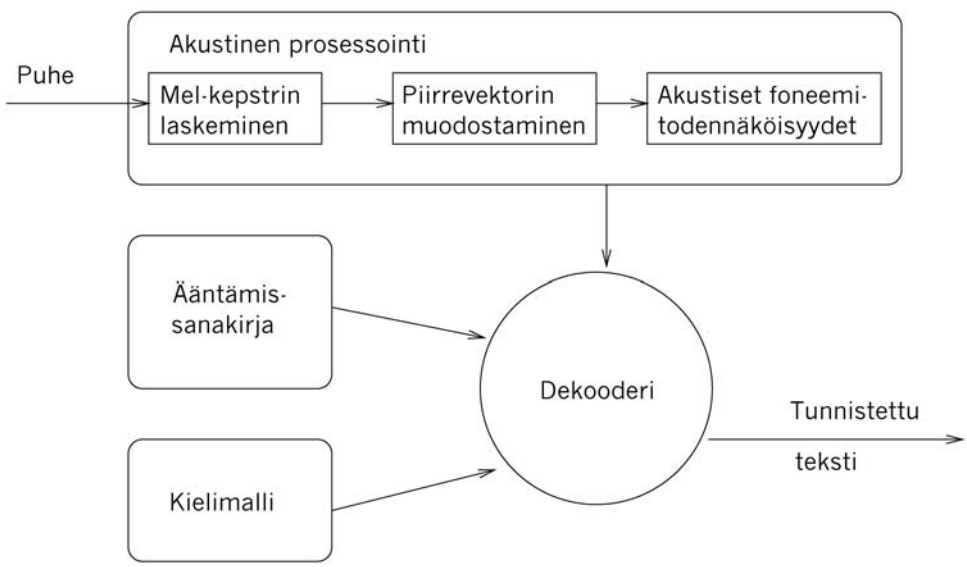
NYKYISTEN PUHEENTUNNISTIMIEN SUORITUSKYKY

Evaluointitehtävien luokittelu

Puheentunnistustehtävien laajasta vaihtelusta johtuen tunnistimen suorituskykyä 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 olevia kielen vaikeuden luokitusperusteita ovat sanaston koko ja kielen syntaksin rajoittavuus. Esimerkiksi jonkin yksinkertaisen kodinkoneen ohjaamiseen riittää hyvin pieni sanasto ja rajallinen kielioppi kun taas toisessa ääripäässä on rajoittamattoman sanaston jatkuva puhe. Tehtävät jaetaan luokkiin myös akustisten ominaisuuksien, kuten kohinan, häiriöiden ja olosuhteiden stabiilisuuden mukaan.

Laajan sanaston tunnistimien suorituskyvystä

Parhaiden nykyisten (englanninkielisten) puheentunnistimien tunnistustarkkuudeksi on mitattu esimerkiksi tavallisille radion ja television uutislähetyksille keskimäärin jopa alle 10 % sanavirheitä (NIST, 1999). Sanavirheet tarkoittavat koko puhelähetyksen tunnistustuloksen sovitusta todelliseen tekstiin niin, että virheiksi lasketaan hävinneet, ylimääräiset ja vaihtuneet sanat. Virheprosentti on siten havaittujen virheiden määrä suhteessa todellisen tekstin sanamäärään. Tunnistustulokselle



Kuva 1. Puheentunnistusjärjestelmän toimintakaavio.

on tyypillistä, että monet puheosuudet ovat lähes täysin virheettömiä, kun taas 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 käytetään usein muun muassa silloin kun haetaan tietoa uutisaineistoista, annotoidaan audio- ja videotallenteita ja indeksoidaan suuria audioaineistoja. Täysin vapaassa sanelussa, jossa kaikki tunnistusvirheet on erikseen jälkikäteen korjattava, virhemäärä voi sen sijaan olla liian suuri. Tosin, jos järjestelmän annetaan adaptoitua tietylle puhujalle ja tietyntyyppisille teksteille, tulos on verrattavissa tottumattomaan konekirjoittajaan; esimerkiksi 20–40 sanaa minuutissa sisältäen virheiden korjaukset (PC Magazine, 1999). Suomenkielisessä tunnistuksessa TKK:lla kehitetty jatkuvan puheen tunnistin saavuttaa esimerkiksi radiouutisissa noin 20 % sanavirheen tarkkuuden. Koska suomen kielessä yksi sana saattaa koostua useasta morfeemista ja vastata useampaa englanninkielistä sanaa, esimerkiksi ”kahvin-juoja-lle-kin = also for a coffee drinker”, saavutettu tunnistustarkkuus ei ole kaukana parhaista englanninkielisistä järjestelmistä.

Suppean sanaston tunnistimien suorituskyvystä

Rajoitetuissa erikoistehtävissä, kuten puhelimen äänivalinnassa tai jonkun laitteen ääniohjauksessa tiettyjen komentojen avulla, puheentunnistimien tarkkuus on yleensä huomattavasti parempi. Kun tehtävä rajoittaa tilannekohtaisten sanavaihtoehtojen joukon pieneksi, vaihtoehtojen akustiset erot ovat usein selkeitä, ja tunnistin toimii lähes virheettömästi. Tällöin tunnistin voi suoriutua tehtävästään riittävän hyvin jopa lievästi häiriöalttiissa olosuhteissa ja suurelle joukolle eri puhujia. Näitä ominaisuuksia on menestyksellisesti hyödynnetty useissa sovelluksissa, joissa yksinkertaisia puhelinpalveluita, kuten, esimerkiksi aikatauluneuvontaa, paikallisia säätiedotuksia, urheilutulosten välitystä ja puhelinluettelopalveluita on automatisoitu. Avainasemassa on tällöin älykäs käyttöliittymä, joka sopivalla tavalla ohjaa käyttäjää kuvaamaan tietotarpeensa puheella niin, että yksittäiset puheentunnistustehtävät, esimerkiksi päivämäärän tai paikannimen lausuminen, jäävät helpoiksi palvelun ollessa silti riittävän joustavaa ja tehokasta. Muita vastaavia tehtäviä, jotka joissakin tapauksissa sopivat hyvin puheen avulla ohjattaviksi, ovat muun muassa erilaiset vammaisten apuvälineet, handsfree-puhelimet ja erilaiset pienet vaatteissa kannettavat laitteet, kuten audio- (ja video-) tallentimet ja soittimet sekä mittalaitteet. Ongelmia tunnistukselle aiheuttavat meluisat käyttötilanteet, epäselvästi tai poikkeavasti puhuvat käyttäjät sekä ei-äidinkielinen puhe ja vieraskieliset nimet. Manuaalinen käyttö sopiikin puheentunnistusta paremmin erityisesti kriittisiin sovelluksiin, kuten pankkipalveluihin tai lentokoneen ohjausjärjestelmiin, joissa edellytetään sataprosenttista toimintavarmuutta kaikissa tilanteissa.

ALAN TUTKIMUS SUOMESSA

Poimintoja tutkimuksen historiasta ja nykytilanteesta

TKK:n Informaatiotekniikan laboratoriossa 1970-luvun lopussa rakennettu automaattinen puheentunnistusjärjestelmä, joka käsitti muutamia tuhansia sanoja, lienee ensimmäinen suomenkielinen puheentunnistin. Tunnistin perustui oppivaan aliavaruusmenetelmään ja redundanttiin hash-osoitukseen. Tämän jälkeen puheentunnistus on ollut laboratorion (nykyään nimeltään tietojenkäsittelytieteen

laitos, jossa toimii adaptiivisen informatiikan tutkimusyksikkö) eräs tärkeimmistä uusien hahmontunnistusalgoritmien testipenkeistä, joilla algoritmien soveltuvuutta hankalan mittausdatan analyysissä on voitu mitata ja verrata *state-of-the-art*-menetelmiin. Samalla saavutetut hyvät tulokset ovat jopa maailmanlaajuisesti vaikuttaneet sekä puheentunnistuksen että itse algoritmien kehityssuuntiin. Tunnetuimpia puheentunnistukseen vaikuttaneita TKK:lla kehitettyjä algoritmeja ovat itseorganisoiva kartta (Kohonen, 1981) ja oppiva vektorikvantisaatio (Kohonen, 1986), jotka kehitettiin akateemikko Teuvo Kohosen johdolla. Puheentunnistuksen merkkipaaluja ovat puolestaan olleet muun muassa foneettinen kirjoituskone (Kohonen, 1988) ja diskreetteihin kätkettyihin Markov-malleihin (HMM) perustuva rajattoman sanaston tunnistin (Torkkola ym., 1991).

Nykyään puheentunnistustutkimus on tärkeällä sijalla sekä itsenäisenä tutkimusalana että monen muun puheen- ja kielentutkimuksen alan tarvitsemana työkaluna (Toivanen & Miettinen, 2001). Tästä esimerkkeinä ovat muun muassa kielentutkimuksen, puhekäännöksen sekä tiedonhakututkimuksen käyttämien puhetallenteiden muuntaminen teksti ksi ja puheanalyysin ja -synteesin tarvitsema puheen segmentointi foneemeiksi. Lisäksi puheentunnistus merkittävänä hahmontunnistusongelmana on noussut maailmanlaajuisesti tärkeäksi hahmontunnistusalgoritmien benchmark-testausvälineeksi, jolla monien uusien algoritmien suorituskykyä voidaan evaluoida suhteessa nykyaikaisiin *state-of-the-art*-menetelmiin.

Puheen signaalinkäsittely

Piirteiden irrotus puhesignaalista on kaikissa *state-of-the-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 piirreirrotusmenetelmiä puheentunnistukseen tutkitaan tällä hetkellä aktiivisesti TKK:n adaptiivisen informatiikan tutkimusyksikön lisäksi TKK:n akustiikan laboratoriossa.

Kohinansieto ja monikielisyys

Automaattisten menetelmien soveltuvuus käytännön puhetilanteisiin, joissa esiintyy monenlaisia akustisia häiriöitä ja kohinaa, on viimeaikoina herättänyt yhä enemmän huomiota puheentunnistustutkimuksessa. Erityisesti Nokian tutkimuskeskuksessa on kehitetty robustia puheteknologiaa, jolla puhutut komennot voidaan tunnistaa ja ymmärtää oikein. Tämäntyyppiset puheentunnistustehtävät ovat suhteellisen kieliriippumattomia, jopa niin, että voidaan kehittää myös monikielisiä tunnistusjärjestelmiä. TTY:n signaalinkäsittelyn laitoksella tutkitaan muun muassa puheen ja äänten separointia taustasta ja TKK:n adaptiivisen informatiikan tutkimusyksikössä puheentunnistusta kohinaisessa ympäristössä.

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.

Kielimallit

Puheentunnistuksen laajempi käyttö tiedonvälityksessä edellyttää yksittäisiä sanoja tai komentolauseita laajempien kokonaisuuksien tunnistamista. Näissä tunnistustehtävissä korostuu kielen mallinnus eli morfologisten, syntaktisten ja semanttisten riippuvuuksien huomiointi, koska sanat usein esiintyvät eri muodoissaan ja ovat riippuvaisia kontekstista (ks. esim. Hirsimäki ym., 2006). 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 puheen sisällöstä esittää todennäköisimpiä hypoteeseja, joita sitten verrataan mitattuun puhesignaaliin. TKK:n adaptiivisen informatiikan 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.

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 oppivien laskentamenetelmien sovelluksena, on tutkimuskohteena TKK:n adaptiivisen informatiikan tutkimusyksikössä. Nykyään tutkimuskohteena ovat myös yleisemmät dynaamiset tilamallit aikasarjoille, joiden avulla HMM:n rajoituksista voidaan päästä eroon.

Multimodaalinen puheentunnistus

Puhe on myös multimodaalinen ilmiö, ja ihmisen kyky integroida kuultu ja nähty puhe sekä muut havainnot puhujasta liittyy puheentunnistustutkimukseen. Multimodaalinen puheentunnistus korostuu erityisesti tilanteissa, joissa esiintyy voimakkaita akustisia häiriöitä tai kuuloaistimus on muuten viallinen. TKK:n laskennallisen tekniikan laboratoriossa tutkitaan visuaalisen puheen havaitsemista ja käsittelyä ja TKK:n adaptiivisen informatiikan tutkimusyksikössä multimodaalista puheentunnistusta, jossa äänen lisäksi tutkitaan puhujan katseen suuntaa ja kohdetta.

Kaupalliset puheentunnistustuotteet

Puheentunnistusteknologiaa on jo Suomessakin tuotu kaupallisten sovellusten avulla osaksi tavallisten ihmisten arkipäivää. Tästä ovat esimerkkeinä muun muassa Nokian matkapuhelimien äänivalinnat, Philipsin puheentunnistusohjelma PC:lle ja Soneran ja Elisan kehittämät automaattiset puhelinpalvelut, kuten numerotiedustelu ja paikallissää. Myös pienet teknologiayritykset ovat panostamassa voimakkaasti suomenkielisten puheentunnistustuotteiden kehittämiseen kuten puhelinluettelo- ja aikataulupalvelut sekä erityisalojen sanelu- ja käännöstehtävät.

Puheen indeksointi ja haku

Laajojen puheaineistojen tunnistustuloksia voidaan käyttää menestyksellisesti myös aineistojen indeksointiin ja siihen perustuvaan tiedonhakuun. Tästä ovat hyvinä esimerkkei-

nä amerikkalaiset SpeechBot- ja SpeechFind-järjestelmät, joilla internetin välityksellä voidaan hakea kiinnostava puheäänite valtavista arkistoista, jotka sisältävät jopa kymmeniä tuhansia tunteja materiaalia, kuten 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 adaptiivisen informatiikan tutkimusyksikössä.

LOPUKSI

Tässä artikkelissa luotiin lyhyt katsaus automaattisen puheentunnistuksen ongelmiin ja esiteltiin nykyaikaisten puheentunnistimien toimintaperiaatetta, matemaattista taustaa ja suorituskykyä sekä alan tutkimusta Suomessa. Puheentunnistustutkimus on tällä hetkellä suuren kiinnostuksen kohteena sekä Suomessa että maailmalla ja menetelmiä kehitellään jatkuvasti yhä uusien kielien ja haastavampien tunnistustehtävien ratkaisemiseksi. Erityisen kiinnostuksen kohteena ovat nykyisin suurten puheaineistojen, kuten radio- ja televisio-ohjelmien, videoiden ja kokoustallenteiden tehokas käsittely sekä toisaalta pienten kannettavien laitteiden puhekäyttöliittymät. Yleistä kieliriippumatonta ratkaisua sanastoltaan rajoittamattoman ja vapaan puheen tunnistamiseen tuskin on lähivuosina tulossa, mutta yhä käyttökelpoisempia ratkaisuja edellä mainittuihin nykysovelluksiin on kuitenkin odotettavissa.

LÄHTEET

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

Jelinek, F., & Mercer, R. L. (1980). Interpolated estimation of Markov source parameters from sparse data. *Proceedings of an International Workshop on Pattern Recognition in Practice.* Amsterdam: North-Holland.

Kohonen, T. (1981). Automatic formation of topological maps of patterns in a selforganizing system. Teoksessa E. Oja & O. Simula (toim.), *Proceeding of the Second Scandinavian Conference on Image Analysis* (2SCIA) (s. 214–220). Helsinki: Suomen Hahmontunnistustutkimuksen Seura.

Kohonen, T. (1986). Learning vector quantization for pattern recognition. Report, TKKF-A601. Espoo, Helsinki University of Technology.

Kohonen, T. (1988). The 'neural' phonetic typewriter. *Computer*, 21, 11–22.

NIST (1999). *Proceedings of DARPA Broadcast News Workshop*. National Institute of Standards and Technology.

PC Magazine. December 1999.

Toivanen, J., & Miettinen, M. (2001). *Puheentutkimuksen resurssit Suomessa.* CSC- Tieteellinen laskenta Oy. http://www.csc.fi/csc/julkaisut/oppaat/pdfs/puhetutk.pdf

Torkkola, K., Kangas, J., Utela, P., Kaski, S., Kokkonen, M., Kurimo, M., & Kohonen, T. (1991). Status report of the Finnish phonetic typewriter project. Teoksessa T. Kohonen, K. Mäkisara, O. Simula, & J. Kangas (toim.), *Artificial Neural Networks, Vol I,* (s. 771–776). Amsterdam: North-Holland.

Hirsimäki, T., Creutz, M., Siivola, V., Kurimo, M., Virpioja, S., & Pylkkönen, J. (2006). Unlimited vocabulary speech recognition with morph language models applied to Finnish. *Computer Speech & Language,* 20, 515–541.

AUTOMATIC SPEECH RECOGNITION

Mikko Kurimo, Adaptive Informatics Research Centre, Department of Information and Computer Science, Helsinki University of Technology

Automatic speech recognition is a significant application of statistical and learning pattern recognition methods. It has a special position as well, because of the amount of applications that are close to even ordinary people. Even though speech recognition is easy for human, the richness and complexity of voice signals makes the task very difficult. In this article the operation, mathematical foundations and practical performance of modern speech recognition systems are briefly presented. Additionally, there is a brief overview of speech recognition research in Finland.

Keywords: Speech recognition.