Puheentunnistus

Mikko Kurimo

Mitä automaattinen puheentunnistus on?

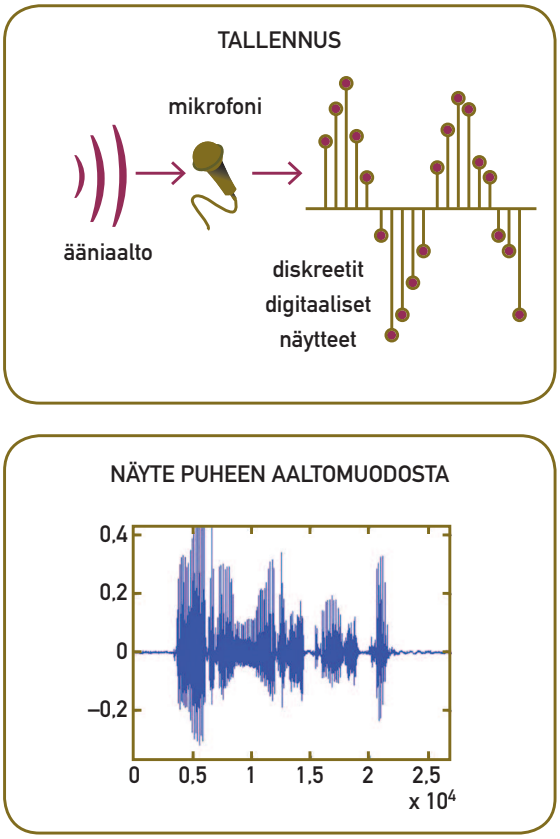
Automaattinen puheentunnistin on laite, joka määrittää ja tulostaa sanan tai tekstin, joka parhaiten vastaa äänitettyä puhesignaalia. Tunnistus perustuu siihen, että ensin äänisignaalista lasketaan puheen äänteitä kuvaavat ominaispiirteet, ja sitten verrataan niitä suuresta puheaineistosta laskettuihin äänteiden tilastollisiin malleihin. Lisäksi tunnistin käyttää suuresta tekstiaineistosta opetettuja tilastollisia sanasto- ja kielimalleja valitakseen tasavahvoista vaihtoehdoista sellaisia sanoja ja tekstejä, joita kielessä todennäköisimmin esiintyy. Tunnistustarkkuutta voidaan usein merkittävästi parantaa rajoittamalla käytettävä sanasto sovelluksen mukaan, esimerkiksi puhelimen äänivalinnassa.

Muita puheenkäsittelyn aloja, joita usein virheellisesti sanotaan puheentunnistukseksi, ovat puhujantunnistus, puheen ymmärtäminen, puheen erottaminen muista tallennetuista äänistä ja puhesynteesi. Niissä on kyllä mahdollista käyttää samankaltaisia tilastollisia menetelmiä, ja niitä myös usein yhdistetään puheentunnistuksen kanssa samoihin sovelluksiin, mutta tehtävä on yleensä jokin muu kuin puheen muuntaminen tekstiksi.

Automaattinen puheentunnistus tai lyhyemmin puheentunnistus, ja usein lyhyesti ASR (engl. **Automatic Speech Recognition**) on tärkeä informaatiotekniikan tutkimus- ja sovellusala. Tämä monitieteinen tutkimusala vaatii asiantuntemusta tietojenkäsittelytieteestä, signaalinkäsittelystä, akustiikasta, fonetiikasta ja kielitieteestä. Sovelluksia löytyy arkipäiväisestä elämästä kuten tiedonhaku automatisoiduista puhelinpalveluista ja tekstin sanelu puhelimelle tai tietokoneelle. Myös kuulo-, näkö- ja liikuntavammaisille on useita palveluja, joissa puhe muutetaan tekstiksi tai komennoiksi, ja joilla voidaan ohjata erilaisia laitteita.

Puheen äänittäminen ja käsittely tunnistusta varten

Puheentunnistuksessa ensimmäinen tehtävä on ilmassa aaltoliikkeenä etenevän äänen tallennus ja muokkaaminen digitaaliseen muotoon. Haasteena on erottaa tutkittava puhe muista ympäristön äänistä kuten tuulettimien ja auton moottorin hurinasta, liikkumisesta syntyvistä äänistä, muista suun tuottamista äänistä ja varsinkin taustalla kuuluvasta toisten ihmisten puheesta. Ihmisaivot ja -korvat ovat olosuhteiden pakosta kehittyneet niin taitaviksi, että tehtävän vaikeutta voi olla hankala edes ymmärtää – varsinkin kun ihminen kykenee ymmärtämään puhetta jopa varsin heikkolaatuisista tallenteista.



Kuva 1. Puhesignaalin käsittely tunnistusta varten.

Tallennuksen ongelmana on se, ettei tallennusta voida rajata yleisesti käytettävillä mikrofoneilla haluttuun puheeseen juuri muuten kuin sijoittamalla mikrofoni mahdollisimman lähelle puhujaa ja soveltamalla vain joitakin hyvin yksinkertaisia ja yleispäteviä menetelmiä taustakohinan vaimennukseen. Puheentunnistimessa mikrofonilla talletettu ja digitoitu puhe jaetaan tarkempaa analysointia varten ensin hyvin lyhyiksi osittain limittäisiksi paloiksi, joiden pituus on tyypillisesti vain kymmenkunta millisekuntia. Sitten jokaisesta palasta eli ikkunasta lasketaan taajuusspektri. Tarkoitus on, että ikkuna on toisaalta niin lyhyt, että sen aikana puheen taajuussisältö ei ehdi muuttua, mutta toisaalta niin pitkä, että spektri voidaan silti luotettavasti laskea. Tarkemmassa analyysissä tutkitaan sitten spektrin tunnistuksen kannalta tärkeimpiä osia eli niitä **tehospektrin** huippuja, jotka sattuvat puheen kannalta oleellisimmille taajuuskaistoille. Tavoitteena on poimia kustakin ikkunasta puheen eri äänteitä parhaiten kuvaavat piirteet niin, että kaikki tunnistuksen kannalta ylimääräinen informaatio, kuten puhujan äänenkorkeus, painotukset ja ympäristön äänet, karsiutuu pois.

Puhesignaalin kuvaaminen ja mallintaminen kompaktina piirrevektorina

Puheen piirteiden analyysin tuloksena muodostetaan jokaisesta analysoidusta puheen palasesta eli ikkunoidusta signaalin pätkästä yksi **piirrevektori**. Piirrevektori sisältää parikymmentä lukuarvoa, jotka on valittu kuvaamaan ikkunassa esiintyvä puheentunnistuksen kannalta merkittävä sisältö mahdollisimman kompaktissa muodossa. Kompakti esitysmuoto on tärkeää sekä jatkossa seuraavan laskentatyön että tilastollisten mallien parametrien määrän ja dimension minimoimiseksi. Hyvä piirre-esitys puhesignaalista

– erottelee äänteet toisistaan
– karsii pois ympäristön häiriöäänet
– on nopea laskea
– on mahdollisimman kompakti.

Tilastollinen malli äänteille muodostetaan määrittämällä kullekin äänteelle todennäköisyysjakauma, joka kuvaa piirrevektorien esiintymistä äännettä vastaavassa tallenteen osassa. Tavallisesti jakauma mallinnetaan moniulotteisella normaalijakaumalla, jossa jokaiselle piirrevektorin alkiolle on suuren puheaineiston perusteella estimoitu keskiarvo ja keskihajonta. Tätä yhdelle äänteelle laskettua todennäköisyysjakaumaa voidaan sitten käyttää laskemaan se todennäköisyys, millä puheesta laskettu piirrevektori voisi olla peräisin tästä jakaumasta. Käytännössä jakaumat ovat niin monimutkaisia, että tarkempaan mallintamiseen

käytetään usein usean normaalijakauman painotettua summaa. Tällaista jakaumamallia kutsutaan **Gaussin mikstuurimalliksi** (GMM eli Gaussian Mixture Model).

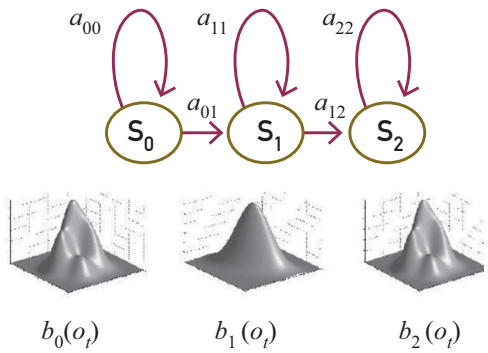
Äänteiden tilastollinen mallintaminen

Puheen rakenneosien eli äänteiden mallintamisessa puheentunnistusta varten ensimmäinen tehtävä on luonnollisesti selvittää, mitä äänteitä puheessa käytetään. Yksi vaihtoehto on valita äänteet puheentutkijoiden kullekin kielelle määrittelemän foneemijoukon eli pienimpien itsenäisten puheen rakenneosien perusteella. Useimmissa kielissä on kuitenkin esitetty vaihtoehtoisia ja erikokoisia foneemijoukkoja. Puheentunnistuksessa on lisäksi huomattu, että monet foneemit voivat esiintyä kontekstista riippuen niin erilaisina variantteina, että niille kannattaa kontekstin perusteella tehdä useampia eri malleja. Muutenkaan parhaan tunnistustuloksen saavuttamiseksi ei ole välttämätöntä noudattaa foneetikkojen esittämää foneemijakoa, vaan esimerkiksi hyvin harvinaiset foneemit kannattaa yhdistää suuremmiksi malleiksi. Niille ei harvinaisuuden vuoksi ole aina määritettävissä riittävän luotettavaa tilastollista mallia.

Koska yhtä mallinnettavaa äännettä vastaavan tallennetun äänisignaalin osan piirteet ovat usein erilaisia äänteen alussa, keskivaiheilla ja lopussa, on tämä otettava huomioon myös äänteille rakennettavassa piirteiden tilastollisessa jakaumamallissa. Äännemalli rakennetaankin tavallisesti useammasta, esimerkiksi kolmesta, peräkkäisestä tilasta, joilla jokaisella on oma jakaumamallinsa ja kestomallinsa. Tällaisen äännemallin etu on joustavuus, koska tilamalli voidaan automaattisesti sovittaa ääninäytteisiin ja laskea kukin jakaumamalli opetusaineistosta tämän sovituksen mukaan. Äännemallia voidaan havainnollistaa ajattelemalla sitä laitteena, jolla on muutama erilainen toimintatila, jotka voidaan erottaa toisistaan epäsuorasti analysoimalla laitteen synnyttämän äänen muutoksia. Tällaista tilamallia

GMM-HMM JÄRJESTELMÄ

– Jokainen tila tuottaa ääniä GMM-mallinsa (Gaussian Mixture Model $b(o)$) mukaan.
– Tätä generoivaa mallia voidaan käyttää myös tekstin muuttamiseen puheeksi.
– Mitä korkeampi a_{ii}, sitä pidempi kesto.



Kuva 2. Tilamalli ja jakaumamalli

kutsutaan **kätketyksi Markov-malliksi** (HMM eli Hidden Markov Model). Nimi kuvaa sitä, että tilan vaihtuminen ei ole suoraan havaittavissa (ts. se on kätketty) ja tilan ominaisuudet, kuten jakaumamalli ja kesto, oletetaan riippumattomiksi aiemmista ja seuraavista tiloista (ns. Markov-oletus).

Yllä kuvatun tilastollisen **tilamallin** (HMM) ja sen sisältämien **jakauma- ja äännekestomallien** perusteella voidaan sovittaa äännenäytteitä tekstiin ja – mikä vielä hyödyllisempää – tunnistaa kokonaisia uusia sanoja ja lauseita. Liittämällä eri äänteiden tilamalleja peräkkäin pitkäksi tilaketjuksi tilastollinen malli voidaan esittää mistä hyvänsä sanasta tai sanajonosta. Vertailemalla tilaketjun mallia uudesta puhenäytteestä laskettujen peräkkäisten piirrevektoreiden jonoon voidaan laskea todennäköisyys, jolla malli voisi generoida tämän näytteen.

Lisäksi vertailemalla eri mallien todennäköisyyksiä voidaan määrittää todennäköisin malli ja

sen avulla muuntaa äänisignaali tekstiksi. Suuresta puhenäytteiden kokoelmasta voidaan lisäksi opettaa tähän kokoelmaan parhaiten sopivat uudet mallit. Tarkempi kuvaus tähän kehitettyjen laskennallisten algoritmien toiminnasta on viitteessä [1].

Kielen mallintaminen ääntämissanakirjana ja tilastollisina kielimalleina

Vaikka pelkkien äännemallienkin avulla voitaisiin teoriassa tunnistaa puhetta muuttamalla äänisignaali äännejonoksi ja jälkikäsittelemällä äännejono edelleen sanoiksi, tämä on kuitenkin osoittautunut käytännössä huonoksi ja äärimmäisen virheherkäksi lähestymistavaksi. Järkevämpää onkin tehdä oletus, että puhe sisältää merkityksellisiä sanoja ja lauseita jollain kielellä. Sen jälkeen laaditaan tästä kielestä suurten tekstiaineistojen perusteella tilastollinen malli ja otetaan se huomioon jo siinä vaiheessa kun signaalia sovitetaan äännemalleihin.

Koska kieli koostuu sanoista, voidaan tilastollisen kielimallin rakennus aloittaa laatimalla sanasto eli leksikko ja määrittämällä kunkin sanan esiintymistodennäköisyys ja todennäköisin ääntämismalli edellisessä luvussa rakennettujen äännemallien jonona. Mikäli joillakin sanoilla on useita tavallisia ääntämistapoja, ääntämissanakirjassa niiden yhteyteen voidaan lisätä kunkin ääntämistavan todennäköisyys. Koska ääntämissanakirjan koko lisää kuitenkin virhetunnistusten määrää, kannattaa useimmiten kuitenkin valita vain yleisin ääntämistapa – varsinkin jos ääntämistavat ovat melko samanlaisia. On huomattava, että kielessä on usein myös sanoja, joita vastaava äännejono on täsmälleen sama, vaikka kirjoitusasu voi olla erilainenkin. Tällaisten sanojen tunnistus on mahdollista vain ympäröivien sanojen eli kontekstin perusteella.

Monissa kielissä puheentunnistuksen tarvitsema sanasto voi olla hyvin suuri, koska on otettava huomioon kaikki mahdolliset taivutusmuodot. Erityisen ongelmallisia tässä suhteessa ovat muun muassa suomensukuiset kielet, joissa taivutusmuodot, yhdyssanat ja erilaiset päätteet ja etuliitteet esiintyvät yleisesti muodostaen usein pitkiä sanoja. Koska pitkät sanat rakentuvat yleensä pienemmistä merkityksellisistä yksiköistä eli morfeemeista, on kuitenkin mahdollista rakentaa leksikko ja kielimalli näiden perusteella. Silloin tarvittavien rakenneosien määrä pysyy järkevän kokoisena.

Koska sanat ja morfeemit eivät kielessä koskaan esiinny satunnaisessa järjestyksessä, voidaan tunnistimen tehtävää merkittävästi helpottaa opettamalla saatavissa olevien tekstiaineistojen perusteella tilastollinen malli sanojen välisistä riippuvuuksista. Käyttämällä tämän kielimallin eri sanajonoille antamia prioritodennäköisyyksiä voidaan tunnistuksessa ratkaisevasti pienentää läpikäytävien vaihtoehtojen määrää ja erotella toisistaan ns. homonyymit eli samalta kuulostavat sanat ja sanayhdistelmät. Rajoittamattomassa sanelussa tyypillisin kielimalli on ns. n-gram-malli, jossa kaikkien sanajonojen todennäköisyydet voidaan laskea yhdistämällä toisiinsa mallin tuntemat lyhyemmät sanajonot ja niiden esiintymistodennäköisyydet. Tarkempi kuvaus n-grammeista ja muiden vastaavien kielimallien toiminnasta on esimerkiksi viitteessä [2].

Puheen muuttaminen tekstiksi äänne- ja kielimallien perusteella

Puheentunnistustehtävä voidaan matemaattisesti määritellä siten, että on löydettävä todennäköisin sanajono puhesignaalin ja käytettävissä olevien äänne- ja kielimallien perusteella. Tästä on edelleen johdettavissa tilamallien hakutehtävä, jonka tulos on paras mahdollinen tilajono äännemallien läpi. Tilajonoa vastaavan sana- tai morfeemijonon prioritodennäköisyys voidaan ottaa haussa huomioon kielimallin avulla. Vaikka käytössä on tehokkaita hakualgoritmeja, niin suora haku on silti jatkuvan puheen tunnistuksessa käytännössä mahdotonta. Sanavälejä ei voida riittävän luotettavasti erottaa esikäsittelyllä, vaan haun täytyy tut-

kia joka kohdassa signaalia mahdollisuus uuden sanan alkamiseen. Tämä seikka yhdessä suuren sanaston kanssa johtaa helposti hakuavaruuden räjähtämiseen. Puheentunnistus on kuitenkin mahdollista jopa reaaliajassa, jos käytetään älykkäitä hakualgoritmeja, jotka osaavat mahdollisimman aikaisessa vaiheessa poistaa kaikki epätodennäköiset sanahypoteesit ja välttävät tehokkaasti kaikkea turhaa laskentaa.

Samalla kun tunnistin etsii parhaan tunnistushypoteesin, voidaan vähällä lisävaivalla tuottaa myös lista seuraavaksi parhaista hypoteeseista (N-best list) tai kompakti sanakaavio (word lattice) josta on helposti erotettavissa suuri joukko parhaita hypoteeseja. Puheentunnistuksen sovellusten lisäksi nämä laajennetut hakutulokset ovat hyödyksi silloin, kun on olemassa tarkempia akustisia ja kielimalleja, joiden käyttäminen alkuperäisessä täydessä haussa on niiden koon takia mahdotonta. Sen sijaan laajennettujen hakutulosten perusteella voidaan parhaat hypoteesit järjestää erittäin nopeasti uudelleen vaikka mallit olisivat suuriakin. Tällöin on mahdollisuus saada tarkennettu tunnistustulos. Puheentunnistuksen hakualgoritmeja ja dekooderiratkaisuja on esitelty tarkemmin viitteessä [3].

Puheentunnistuksen sovellukset, suorituskyky ja haasteet

Sovelluksista yleisesti

Puheentunnistuksen sovelluksista ensimmäisenä tulee monelle mieleen muiden teknisten laitteiden ohjaaminen ilman käsin tapahtuvaa yksityiskohtaista nappien painelua. Tähän liittyy usein myös se taustatoive, ettei ihmisen tarvitsisi tarkalleen kuvata haluttua toimintaa, kuten mitä nappeja painetaan ja miten vipuja käännetään, vaan kone osaisi samalla älykkäästi päätellä, mitä ihminen oikeastaan haluaa ja toimia sen mukaan. Usein on kuitenkin tuloksen kannalta parempi, että sivuutetaan puheohjaus ja ihmisen toiveiden arvailu ja ohjataan laitetta suoraan, esimerkiksi käyttämällä itse tietokonetta tai ohjaamalla autoa. Puheohjaus on kuitenkin jo nykyään hyödyllinen sovellus silloin, kun on suoritettava melko yksinkertaisia tehtäviä ilman käsien apua tai esimerkiksi puhelimen välityksellä. Erityisesti kuulo-, näkö- ja liikuntavammaisten toimintaa helpottavat palvelut, joissa puhe muutetaan tekstiksi tai komennoiksi, joilla voidaan ohjata erilaisia laitteita.

Puheentunnistimen käyttöön liittyy helposti varsin epärealistisia odotuksia. Jos laitteelle voidaan puhua kuten ihmiselle, oletetaan helposti, että se on muullakin tavoin älykäs ja kykenee suorittamaan vaativia tehtäviä ikään kuin palvelija, jolle ei tarvitse yksityiskohtaisesti määritellä mitä tehdään. Koska ihminen on kehittynyt erittäin taitavaksi erottamaan toisen ihmisen puheen muista häiriöäänistä ja jopa usean päällekkäisen puhujan joukosta, toivotaan, että koneellinenkin puheentunnistus kykenee samaan. Samoin on laita myös tunnistuksen virheettömyydessä puhetapaa, -tyyliä ja -nopeutta muunneltaessa sekä eri puhujien välisten erojen huomioinnissa. Vaikka nykyaikainen puheentunnistin kykeneekin tunnistamaan sen äänne- ja kielimalleille opetettua aineistoa vastaavaa selkeää puhetta jopa hämmästyttävän hyvin, voi puhetilanteen ja -tavan muuttuminen johtaa tunnistustarkkuuden romahtamiseen. Lisäksi on huomattava, että tunnistimen signaalista käyttämä informaatio – vain äänteiden erotteluun optimoidut lyhytaikaisen spektrin piirteet – kattaa hyvin pienen osan siitä kaikesta oheisinformaatiosta, jota ihminen usein huomaamattaankin käyttää epäselvän puheen ymmärtämiseksi. Tämä oheisinformaatio liittyy äänensävyyn, puhujan muihin ominaisuuksiin, puheenaiheeseen, kuulijan aiempiin kokemuksiin ja monenlaiseen näköaistin välityksellä tilanteesta saatavaan tietoon.

Puhekäyttöliittymät

Puheentunnistuksen sovellukset on tuloksille asetettavien vaatimusten perusteella jaettavissa kolmeen pääryhmään: käyttöliittymät, sanelu ja

tiedonhaku. Käyttöliittymissä tunnistetaan puhuttuja komentoja ja tuloksena on joko visuaalinen tai synteettinen vaste tai suoraan ohjattavan laitteen toiminta (ks. s. XX–XX Turunen). Esimerkkejä laitteista, joissa puheentunnistusta voidaan käyttää käyttöliittymänä, ovat matkapuhelin, tietokone, auton oheislaitteet (kuten navigaattori) ja automatisoidut tietopalvelut (kuten numerotiedustelu, paikallissää ja erilaiset varauspalvelut).

Visuaalinen vaste varmistaa tunnistustuloksen näyttämällä valitun komennon laitteen näytöllä ja synteettinen vaste toistaa sen laitteen omalla äänellä. Yleensä tunnistustarkkuuden komentojen osalta on oltava lähes täydellistä, tai muuten käyttäjä joutuu antamaan komennon uudelleen. Tunnistusta helpottaa merkittävästi se, että kussakin käyttötilanteessa on yleensä varsin rajallinen määrä toimintavaihtoehtoja – esimerkiksi soitettaessa puhelimella vain muistissa olevat nimet. Tunnistukselle haasteita asettavat käyttötilanteessa kuuluvat taustaäänet ja se, että käyttäjäryhmä voi olla suuri eivätkä kaikki ole välttämättä kovin harjaantuneita käyttöliittymän toimintaan.

Sanelu

Sanelussa tunnistetaan jatkuvaa puhetta. Sen tuloksena on tunnistettu puhe tekstinä, yleensä näytöllä esitettynä siten, että käyttäjä voi tarvittaessa korjata virheelliset sanat. Tavallisia esimerkkejä sanelusta ovat erilaiset raportit, kuten lääkärin muistiinpanot, sähköpostit ja tekstiviestit. Samantyyppisellä puheentunnistimella voidaan jälkikäteen käsitellä myös erilaisia tallennettuja puheäänitteitä kuten haastatteluja, kokoustilanteita ja puhelimeen jätettyjä ääniviestejä. Kun puhetta muunnetaan tekstiksi, tavoitteena on tarkkuus ja joissakin tapauksissa myös nopeus. Puhetta tallennettaessa tämä tavoite on yleensä tiedossa, joten puhuja voi vaikuttaa siihen puhumalla rauhallisesti ja selkeästi. Lisäksi hän voi hakeutua rauhalliseen ja hiljaiseen ympäristöön ja pitää mikrofonia huolellisesti suunsa edessä.

Sanelun tapainen sovellus on myös puheen kääntäminen toiselle kielelle. Siinä tunnistettu teksti käännetään automaattisilla kielenkäännöstyökaluilla ja lopuksi joko tallennetaan vieraskielisenä tekstinä tai soitetaan puhesyntetisaattorin avulla vieraskielisenä puheena. Koska koneellinen kääntäminen on vähintään yhtä haastava tehtävä kuin puheentunnistus ja hyvin herkkä tunnistusvirheille, sovelluksen aihepiiri on yleensä tiukasti rajattu, esimerkiksi vain tavallisten matkailuun liittyvien fraasien kääntäminen.

Tiedonhaku

Kolmas pääryhmä sovelluksia on tiedonhaku. Siinä tunnistuksen tavoitteena on ns. raakatekstin tuottaminen luokiteltavaksi tai indeksoitavaksi edelleen sisällön perusteella tiedonhakua varten. Tyypillisiä tehtäviä ovat puheiden haku tietyistä aiheista, äänitteiden indeksointi ja selailu sekä yhteenvetojen laatiminen. Tunnistusta helpottaa se, ettei raakatekstin tarvitse olla virheetöntä, vaan relevanssin määrittäminen sisällön kannalta voi toimia riittävän hyvin, vaikka jopa puolet sanoista sisältäisi virheitä. Tunnistuksen haasteena taas on se, että käsiteltävissä äänitteissä puhuja on yleensä suunnannut puheensa koneen sijasta toisille ihmisille. Puhe voi myös olla nopeaa ja epäselvää eikä äänitysympäristön ja mikrofonin sijoittelun huomiointi välttämättä ole tunnistuksen kannalta optimaalinen. Äänitteet voivat olla televisio-ohjelmia, videokuvauksia tai pöytämikrofonilla äänitettyjä haastatteluja tai kokouksia.

Käytännön ongelmia ja ratkaisuja

Opetusaineisto

Koska automaattinen puheentunnistus perustuu puhe- ja tekstiaineistoista opetettuihin tilastollisiin malleihin, on tunnistimen suorituskyky suuresti riippuvainen käytetyn aineiston kattavuudesta ja sopivuudesta. Tyypillisesti puhujariippumatto-

maan tunnistukseen tarvitaan ääninäytteet usealta sadalta eri ihmiseltä. Tekstiaineiston koko on sanastosta ja aiheesta riippumattomassa sanelussa oltava kymmeniä tai jopa satoja miljoonia sanoja. Koska tilastolliset mallit kuvaavat opetusaineiston keskimääräisiä piirteitä, voi normaalista poikkeava puhetapa tai puheenaihe tuottaa ylipääsemättömiä ongelmia. Tyypillisiä esimerkkejä ovat pienen lapsen puhe tai lyhenteitä vilisevä teknillisen erityisalan puhe.

Aineiston annotoinniksi riittää yleensä lausetaso eli puhetta vastaava teksti. Tarkempi annotointi sanojen, äänteiden ja tilamallin tilojen alkamis- ja päättymishetkistä saadaan riittävällä tarkkuudella tehtyä puheentunnistimella. Tämä tekstiä vastaavan puheen segmentointi äänteisiin onkin yksi muiden puheentutkijoiden paljon käyttämä puheentunnistuksen sovellus ja sivutuote. Lisäksi on myös mahdollista opettaa tilastollisia malleja eteenpäin niin, että tekstin sijasta käytetään puheentunnistustuloksia.

Äänne- ja kielimallien adaptointi

Puhujariippumatonta tunnistusta parempaan tulokseen päästään, jos kullekin puhujalle voidaan opettaa oma henkilökohtainen malli. Tähän tarvitaan useampi tunti annotoitua puhetta. Toinen vaihtoehto on käyttää puhuja-adaptaatiota puhujariippumattoman mallin muokkaamiseksi. Tulos ei ole yhtä tarkka, mutta jo muutamalla kymmenellä lauseella päästään selvästi yleismallia parempiin tuloksiin.

Puhuja-adaptaation laskemiseen on kaksi eri tapaa. Joko tilastollisen mallin kutakin parametria muokataan adaptointidataan paremmin sopivaksi tai sitten estimoidaan puheen piirteille yleinen muunnosmatriisi, jolla kaikki tilastolliset mallit saadaan paremmin sopimaan adaptointidataan. Jälkimmäisellä tavalla päästään tulosparannuksiin huomattavasti nopeammin, mutta se ei ota yhtä tarkasti huomioon eri äänteissä tapahtuvia erilaisia puhujakohtaisia korjauksia, koska muunnosmatriiseja ei voida pienellä datamäärällä estimoida äännekohtaisesti. Puhuja-adaptaatio korjaa melko hyvin myös varsinaisesta opetusaineistosta poikkeavia äänitysolosuhteita ja yksinkertaista jatkuvaa taustaääntä kuten moottorin hurinaa.

Jos puheenaihe tai puhetyyli on tiedossa, voidaan tilastollisia kielimalleja adaptoida paremman tunnistustuloksen saavuttamiseksi. Tämä vaatii kuitenkin melko paljon uuteen aiheeseen liittyvää tekstiaineistoa (vähintään sadasta tuhannesta miljoonaan sanaa), ellei samalla voida rajoittaa myös kielen rakennetta. Tämä on kuitenkin käytännössä usein hankalaa. Kielimallien adaptaatiossa opetetaan kullekin puheenaiheelle tai puhetyylille oma tilastollinen mallinsa ja sitten tunnistettavan puheen perusteella laaditaan sopiva kombinaatio yleiskielimallin ja valitun aihemallin välillä. Yleiskielimallin käyttö on usein hyödyllistä koska aihemallin yleistämiskyky on pienemmän opetusaineiston takia heikompi.

Myös murteita tai puhekieltä voidaan puheentunnistimen kannalta pitää erilaisina puhetyyleinä. Adaptointi on kuitenkin vaikeaa, koska yleiskielimallista ei välttämättä ole hyötyä, vaan sekä sanastolle että sanojen välisille riippuvuuksille on opetettava uudet mallit. Myös sanojen ääntäminen voi olla niin erilaista, että tarvitaan uusi ääntämissanakirjakin. Lisäksi riittävän laajan kirjallisen opetusaineiston kerääminen vain puhuttuna esiintyvälle kielelle on työlästä ja usein jopa mahdotonta.

Puheentunnistus meluisassa ympäristössä

Jos puheentunnistusta joudutaan käyttämään huonoissa äänitysoloissa ja meluisassa ympäristössä, voidaan tunnistusta tehostaa joko opettamalla uudet tilanteeseen sopivat mallit tai mallintamalla melu erikseen omalla mallillaan. Huonot äänitysolosuhteet voivat johtua yksinkertaisesti siitä, että mikrofoni ei ole suun lähellä, vaan jossain kauempana. Uudet mallit opetetaan joko keräämällä uutta opetusdataa puheesta kohdeympäristössä, esimerkiksi liikkeellä olevassa autossa, tai muut-

tamalla tilastollisten mallien käyttämiä puheen piirteitä sellaisiksi, että melu suodattuu niistä pois. Erityisen vaikeasti suodatettavaa melua on taustalla kuuluva puhe, koska siinä esiintyy samanlaisia äänteitä kuin tunnistettavassa puheessa. Tunnistimen on vaikea erottaa niitä toisistaan ja tietää, mitä pitäisi tunnistaa ja mitä ei.

Mikäli taustamelu on sopivan kaavamaista ja siitä on riittävästi näytteitä, jotta sille voidaan opettaa oma tilastollinen mallinsa, on mahdollista yhdistää se suoraan puheentunnistusprosessiin. Tällöin tunnistin käyttää yhtäaikaisesti sekä puheen että melun malleja löytääkseen parhaiten niihin molempiin sopivan tunnistustuloksen. Tämä on kuitenkin laskennallisesti erittäin raskasta ja vaatii kaikille erilaisille melutyypeille omat mallinsa.

Puheentunnistus eri kielillä

Toistaiseksi automaattisia puheentunnistimia on kehitelty vain pienelle osalle maailman kielistä. Koska painopiste on ollut valtakielissä (erityisesti englanti), monet puheentunnistimissa käytetyt ratkaisut on kehitetty nimenomaan niitä varten.

Viime aikoina on kuitenkin kiinnitetty yhä enemmän huomiota sellaisiin menetelmiin, jotka ovat helposti siirrettävissä uusiin kieliin ilman suurta manuaalista kehitystyötä. Hankalimmat ongelmat, joita erilaiset kielet aiheuttavat, eivät välttämättä liity erilaisiin äänteisiin, vaan siihen, miten sanat ja lauseet rakentuvat eri tavoin ja erilaisista osista. Nämä ongelmat vaikeuttavat erityisesti sanasto- ja kielimallien kehittämistä. Luonnollisesti suurimmat ongelmat koskevat kieliä, joille ei ole kehittynyt edes vakiintunutta kirjoitustapaa tai joilla tapa kirjoittaa ja puhua kieltä poikkeavat paljon toisistaan.

Kirjallisuus

[1] Lawrence R. Rabiner (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77 (2)*, 257–286.

[2] Joshua T. Goodman (2001). A bit of progress in language modeling. *Computer Speech & Language, 15 (4)*, 403–434.

[3] Xavier L. Aubert (2002). An overview of decoding techniques for large vocabulary continuous speech recognition. *Computer Speech & Language, 16 (1)*, 89–114.