AUTOMAATTINEN PUHEENTUNNISTUS

Teemu Salminen

Opinnäytetyö
Syyskuu 2015
Tietotekniikka
Sulautetut järjestelmät
ja elektroniikka

TAMPEREEN AMMATTIKORKEAKOULU
Tampere University of Applied Sciences

TIIVISTELMÄ

Tampereen ammattikorkeakoulu
Tietotekniikan koulutusohjelma
Sulautetut järjestelmät ja elektroniikka

TEEMU SALMINEN
Automaattinen puheentunnistus

Opinnäytetyö 26 sivua
Syyskuu 2015

Tässä opinnäytetyössä käydään läpi tavanomaisen Markovin piilomalleihin pohjautuvan automaattisen puheentunnistusjärjestelmän toimintaperiaate. Työn tarkoituksena on antaa lukijalle yksinkertaistettu kuva nykyisten käytössä olevien tunnistimien toiminnasta ja tulevaisuuden kehityksen suunnasta.

Automaattinen puheentunnistus tai lyhyesti ASR on merkittävä tilastollisten ja hahmontunnistus menetelmien sovellus, joka mahdollistaa luonnollisen kielen käytön ihmisen ja koneen välisessä vuorovaikutuksessa. Puheentunnistusjärjestelmien ydin koostuu tilastollisilla menetelmillä estimoiduista malleista, jotka edustavat tunnistettavan puhekielen eri rakenneosia eli äänteitä, joita tunnistuksessa verrataan puhesignaalista laskettujen äänteiden ominaispiirteisiin. Markovin piilomallit tarjoavat yksinkertaisen ja tehokkaan tavan käsittelemään puheen ajallista vaihtelevuutta, jonka seurauksena lähes kaikki nykypäivän automaattisen jatkuvan puheentunnistusjärjestelmien tilastolliset äännemallit perustuvat Markovin piilomalleihin.

Tässä työssä tehdään aluksi lyhyt katsaus automaattisen puheentunnistuksen historian päävaiheisiin ja nykytilaan, jonka jälkeen työssä esitellään Markovin piilomalleihin (HMM) pohjautuvan automaattisten puheentunnistusjärjestelmän yleinen rakenne ja toiminta.

Avainsanat: Puheentunnistus, Markovin piilomallit

ABSTRACT

Tampereen ammattikorkeakoulu
Tampere University of Applied Sciences
ICT Engineering
Embedded Systems

TEEMU SALMINEN
Automatic speech recognition

Bachelor's thesis 26 pages
September 2015

This thesis presents the architecture of conventional automatic speech recognition system based on the Hidden Markov Models. The aim of this work is to give the reader a simplified picture of modern recognizers and brief overview of future direction of development.

Automatic speech recognition is a significant application of statistical and learning pattern recognition methods, which allows the use of natural language between man and machine interaction. The core of speech recognition systems consists of a set of statistical models, which represents the various sounds of the recognizable language that during recognition are compared to the computed characteristics of sounds of the speech signal. Hidden Markov models provide a simple and effective way to deal with the temporal variability of the speech, as a consequence, almost all present day automatic continuous speech recognition systems statistical models are based on hidden Markov models.

In this work we will first take a brief overview of the main stages of the history and practical performance of modern speech recognition systems, and then the general architecture and operation of HMM–based speech recognition systems are presented.

Keywords: Speech recognition, hidden Markov models

SISÄLLYSLUETTELO

1 JOHDANTO

Automaattinen jatkuva puheentunnistus on ollut kehityksen alla jo vuosikymmeniä, johtuen sen monista potentiaalisista mahdollisuuksista. Nykyisin puheentunnistustekniikkaa käytetään yleisesti sanelusovelluksissa, jossa puhe taltioidaan tekstimuotoon ja erilaisissa mobiililaitteissa, joissa käyttäjä voi puhekomentojen avulla soittaa, sanella ja lähettää sähköposti- ja tekstiviestejä, tai etsiä tietoa internetistä. Puheentunnistustekniikan kehityksen tavoitteena on luoda älykkäitä koneita, jotka kykenevät kuulemaan ja ymmärtämään puhuttua informaatiota, riippumatta luonnollisen kielen epäselvyydestä ja monimutkaisuudesta. Nykyisin, jos järjestelmä on koulutettu oppimaan yksittäisen henkilön puhesignaalin ominaisuudet, niin laajan sanaston omaavan tunnistimen sanatarkkuus voi saavuttaa hyvissä akustisissa olosuhteissa jopa lähes virheettömän lopputuloksen.

Useimmat nykyisistä tunnistusjärjestelmistä perustuvat tyypillisesti tilastollisiin malleihin, jotka edustavat tunnistettavan kielen eri äänteitä. Lähes kaikkien nykyisten puheentunnistusjärjestelmien tilastolliset äännemallit pohjautuvat Markovin piilomalleihin (HMM). Tässä työssä tarkastellaan näihin tilastollisiin äännemalleihin perustuvan tavanomaisen HMM–pohjaisen puheentunnistusjärjestelmän toimintaperiaatetta kirjallisuuden pohjalta.

Ensimmäisessä osiossa käydään lyhyesti läpi mitä puheentunnistus on ja puheentunnistuksen historian päävaiheet sekä tarkastellaan hieman nykyisten tunnistimien tarkkuutta. Seuraavassa osiossa käydään läpi puheentunnistuksen päävaiheet ja perehdytään puheen piirteidenirrotukseen eli prosessiin, jossa puhesignaalista lasketaan äänteitä kuvaavat ominaispiirteet tunnistusta varten. Tämän jälkeen työssä kuvataan äänteiden tilastollisessa mallinnuksessa käytettyjen Markovin piilomallien (HMMs) toimintaperiaate sekä tutustutaan akustisessa mallinnuksessa käytettyihin erilaisiin äännemalleihin. Tästä jatketaan käymällä läpi kielen mallintaminen, jossa lasketaan todennäköisyydet sanoille ja sanayhdistelmille tilastollisten kielimallien avulla. Lopuksi tutustutaan puheen dekoodaukseen, eli prosessiin jolla puhe muunnetaan tekstiksi äänne- ja kielimallien perusteella.

2 PUHEENTUNNISTUKSEN YLEISKUVAUS JA NYKYTILA

Automaattinen jatkuva puheentunnistus (engl. ACSR eli Automatic Continuous Speech Recognition) voidaan määritellä itsenäisenä, tietokone ohjatulla transkriptiona puhutulle kielelle reaaliajassa. Pähkinänkuoressa ACSR on järjestelmä, joka määrittää ja tulostaa sanan tai tekstin, jonka koulutetut tilastolliset äännemallit parhaiten vastaavat äänitetystä puhesignaalista laskettuja puheen äänteitä kuvaavia ominaispiirteitä.

Puheentunnistus on yksi tekoälytutkimuksen merkittävistä osa-alueista, jonka historian voidaan olettaa alkaneen vuonna 1950, kun Alan Turing julkaisi tekoälytutkimuksen virstanpylväänä pidetyn artikkelin *Computing machinery and intelligence*, jossa Turing määritteli käytännönläheisen kokeen, jolla voisi mitata tietokoneen ihmismäisyyttä. Kaksi vuotta myöhemmin yhdysvaltalainen Bell Labs niminen tutkimusorganisaatio kehitti ensimmäisen puheentunnistusjärjestelmän, nimeltä Audrey (Automatic Digit Recognizer), joka kykeni tunnistamaan ainoastaan yksittäisen henkilön lausumia numeroita 1 ja 9 välillä. Vasta kymmenen vuotta myöhemmin IBM esitteli 1962 maailmannäyttelyssä sen kehittämän ”Shoebox” tunnistimen, joka kykeni ymmärtämään huimat 16 puhuttua sanaa, johon lukeutui numerot nollasta yhdeksään ja aritmeettisten laskutoimituksien äänikomennot. 1970–luvulla puheentunnistuksen kehitys otti merkittäviä edistysaskelia, kun Yhdysvaltain asevoimien tutkimusorganisaatio DARPA aloitti viisi vuotta kestäneen puheentunnistuksen tutkimuksen rahoittamisen. DARPA:n kiinnostuksen ja rahoituksen seurauksena syntyi Carnegie Mellon–yliopiston kehittämä Harpy niminen puheentunnistusjärjestelmä, joka kykeni tunnistamaan yli 1000 sanaa ja saman sanan eri ääntämisen variaatioita. Harpy järjestelmä oli merkittävä edistysaskel, sillä sen kehityksen seurauksena syntyi tehokas heuristinen hakualgoritmi tekniikka, nimeltä beam search. Modernin HMM–pohjaisen jatkuvan puheentunnistusjärjestelmän perusta luotiin 1980–luvulla tilastollisten Markovin piilomallien (HMMs) käyttöönoton johdosta, jotka sanamallien käytön ja niistä yhtäläisyyksien etsimisen sijaan tarkastelevat todennäköisyyksiä, joilla tuntemattomat äännähdykset voisivat olla sanoja. Vuonna 1985, Kurzweil Applied Intelligence julkaisi ensimmäisen speech-to-text ohjelmiston, joka ymmärsi 1000 puhuttua sanaa, ja josta kaksi vuotta myöhemmin julkaistiin päivitetty versio, jonka sanasto kasvoi jopa 20 000 sanaan. Puheentunnistus tekniikka kokonaisuudessaan oli kuitenkin vielä riippuvainen diskreettisestä lausahdus järjestelmästä, joka teki lyhyen tauon pitämisen sanojen välillä tarpeelliseksi. 1990–luvulla useat eri yrityk-

set alkoivat julkaista kaupallisia puheentunnistus ohjelmistoja, joista Dragon Systems julkaisi vuonna 1997 ensimmäisen jatkuvan puheentunnistus ohjelmiston ”Naturally Speaking”, joka kykeni tunnistamaan normaalia jatkuva-aikaista puhetta. (Sadewo, B. 2012)

Nykyisten englanninkielisten laajan sanaston jatkuvan puheentunnistusjärjestelmien tunnistustarkkuudeksi on mitattu muun muassa tavallisille radion ja television uutislähetyksille keskimäärin 20 % sanavirhettä. Sanavirheillä tarkoitetaan koko lähetyksen tunnistustuloksen vertausta varsinaiseen tekstiin siten, että virheelliseksi tunnistustulokseksi lasketaan hävinneet, ylimääräiset ja vaihtuneet sanat. (Kurimo, M. 2008)

Suppean sanaston puheentunnistimissa, jossa tunnistus on rajoitettu tilannekohtaiseen puheeseen, voidaan saavuttaa jopa lähes virheetön tunnistustulos, sillä rajoitetussa sanastossa samalta kuulostavien sanojen määrä jää usein hyvin pieneksi, jolloin sanavaihtoehtojen akustiset erot ovat usein selkeitä ja näin ollen helpommin tunnistettavissa. Tämän tyyppiset tunnistimet voivat suoriutua tehtävästään tarpeeksi hyvin jopa hieman häiriöalttiissa olosuhteissa ja usean erityyppisen puhujan ymmärtämisestä. (Kurimo, M. 2008)

Kaupallisesti saatavilla olevat ASR järjestelmät vaativat yleensä lyhyen ajan käyttäjän puheäänen koulutuksen, jolloin normaalitahtisen jatkuva-aikaisen puheen kaappaus, laajalla sanavarastolla on mahdollista hyvin suurella tarkkuudella. State-of-art puheentunnistusjärjestelmä, jolle on koulutettu yksittäisen henkilön puhesignaali, voi optimaalisissa olosuhteissa saavuttaa jopa 99 % tarkkuuden. Optimaalisilla olosuhteilla tarkoitetaan vähäistä taustamelu ympäristöä ja että käyttäjän puheominaisuudet (esim. aksentti) vastaavat järjestelmälle opetetun sanavaraston puhetallenteiden tietoja.

Vaikka ASR–teknologia ei ole vielä siinä vaiheessa, jossa koneet ymmärtäisivät kaikkea, kenen tahansa henkilön puhetta, tai missä tahansa ääniympäristössä, niin sitä käytetään useassa eri sovelluksessa ja palvelussa. Automaattisen puheentunnistuksen tutkimuksen perimmäinen tavoite on mahdollistaa tietokoneen tunnistaa reaaliajassa, 100 % tarkkuudella kaikki sanat, jota kuka tahansa henkilö on puhunut, riippumatta taustamelusta ja puhujan puhetavasta tai aksentista.

3 HMM–POHJAISEN TUNNISTIMEN RAKENNE

Markovin piilomalleja (HMMs) pidetään parhaimpana menettelytapana nopean ja tarkan puheentunnistusjärjestelmän toteuttamisessa. Useimmat moderneista automaattisista jatkuvan puheentunnistusjärjestelmistä käyttävät jatkuvatiheyksisiä Markovin piilomalleja (CDHMM) käsittelemään puheen ajallista vaihtelevuutta. HMM–pohjaiset laajan sanaston jatkuvan puheentunnistusjärjestelmät (LVCSR) perustuvat ennalta estimoituihin äänteiden akustisiin malleihin, jotka koostuvat tyypillisesti tuhansista parametreista. Tämän lisäksi ne käyttävät apunaan suuria leksikkoja (ns. ääntämissanakirjoja) ja kielimalleja mallintamaan tunnistettavan kielen rakennetta. Monimutkaiset akustiset mallit vaativat kuitenkin mittavaa äännemallien koulutusta, jotka on estimoitava tunnistimen opetusvaiheessa. Esimerkiksi Englannin sanakirjassa olevat ääntämisen symbolit jokaisen sanan vieressä edustaa foneemeja, jossa kukin foneemi on erillinen rakenneosa englanninkielen puheessa. Nämä symbolit kertovat, miten jokainen sana tulisi lausua. Tietokoneilla ei ole tätä luontaista tietoa sanojen ääntämisestä, joten tunnistusjärjestelmille on ensin opetettava miltä kukin foneemi kuulostaa. Opettamalla tunnistimelle englanninkielen jokaisen foneemin, niin kone voi päätellä miltä jokainen sana leksikossa kuulostaa. (Gales, M. & Young, S. 2008; Gmoore 2005)

Aluksi järjestelmälle on annettava joukko puhenauhoituksia sekä niiden oikein tulkintoja, ja määrittää puhesignaalista mikä kukin foneemi on määrittämällä tarkalleen milloin se alkaa ja loppuu. Puheentunnistin voi tämän jälkeen aloittaa opettelu prosessin erilaisten algoritmien avulla, käymällä läpi puhetallenteita ja rakentaa foneemi esimerkkien tietokantaa. Foneemit voidaan määritellä eriävistä energia tasoista eri taajuusalueilla. Analysoimalla foneemien akustisia esimerkkejä, järjestelmä voi selvittää mikä jokaisen foneemin keskimääräinen taajuusrakenne on. Tästä voidaan johtaa tilastollisen malli foneemille, joka ei ainoastaan ilmaise jokaisen foneemin keskiarvoa, mutta myös sen mahdollista variaatiota. Koska nämä keskiarvot johdetaan koulutuksen aikana, eri kouluttajan aksentti vaikuttaa, miten malli edustaa kutakin foneemia. Toisin sanoen, puheentunnistin, joka on koulutettu amerikanenglannin kielellä, voi olla ongelmia ymmärtää vahvalla Lontoon aksentilla puhuttua englanninkieltä. (Gales, M. & Young, S. 2008; Gmoore 2005)

Puheentunnistusjärjestelmät voidaan äännemallien koulutuksen aikana suunnitella joko puhujasta riippumattomaksi tai puhujasta riippuvaiseksi. Puhujasta riippuvaiset järjestelmät on suunniteltu tunnistamaan suurella sanatarkkuudella yksittäisen henkilön puhetta, kun taas puhujasta riippumattomat järjestelmät on kehitetty tunnistamaan tietyn tyyppistä puhetta (esim. Amerikan Englanti). Puhujasta riippumattomien järjestelmien tilastolliset äännemallit koulutetaan usean eri henkilön (vanhusten, nuorten, miesten, naisten jne.) puheesta, jolloin tämän tyyppiset järjestelmät saavuttavat paremman joustavuuden, mutta ei yhtä suurta sanatarkkuutta kuin puhujasta riippuvaiset järjestelmät.

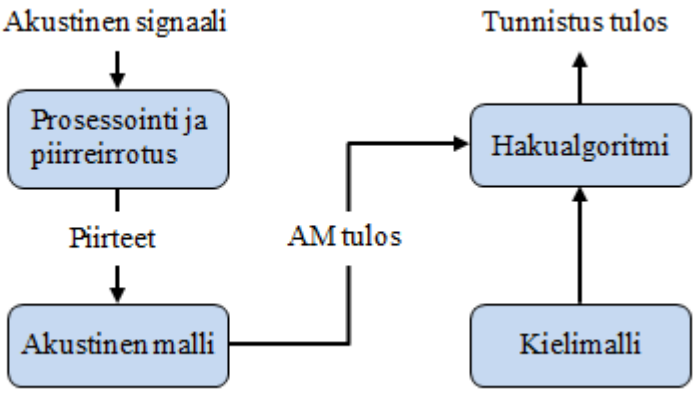
HMM–pohjaisen automaattisen jatkuvan puheentunnistusjärjestelmän (ACSR) toiminta perustuu siihen, että puhesignaalista lasketaan puheen eri äänteitä eli foneemeja kuvaavat ominaispiirteet, jonka jälkeen piirteitä verrataan edellä kuvattuihin isosta puheaineistosta opetettujen foneemien tilastollisiin malleihin. Tämän lisäksi tunnistin käyttää kielenmallintamisen apuna suuresta tekstiaineistosta opetettuja tilastollisia sanasto- ja kielimalleja valitakseen vahvoista vaihtoehdoista sellaisia sanoja, joita kielessä kaikkein todennäköisimmin esiintyy ja jotka ovat yhteensopivia puhutun viestin kontekstin kanssa. Automaattisen puheentunnistusjärjestelmän tyypillinen rakenne voidaan jakaa yleensä seuraavasti eteneviin prosesseihin:

- **Piirrevektorien muodostaminen**: näytteistetty puhesignaali paloitellaan lyhyisiin kehyksiin, joista poimitaan puheen eri äänteitä parhaiten kuvaavat piirteet.
- **Akustinen mallintaminen**: kehysten piirteistä lasketaan todennäköisyys äänteille eli foneemeille; lasketaan foneemisekvenssin sopivuus annetulle puhesignaalille.
- **Kielimallinnus**: määritetään todennäköisyys sanoille- ja sanayhdistelmille; estimoidaan seuraavan sanan todennäköisyys edeltävien sanojen perusteella.
- **Puheen dekoodaus**: haetaan todennäköisin puheen sisältöä vastaava sanajono akustisten- ja kielimallien todennäköisyyksien perusteella. (Kurimo, M. 2008)

Puheentunnistuksessa ensimmäinen tehtävä on mikrofonilla tallennetun analogisen puhesignaalin muokkaaminen digitaaliseen muotoon. Tallennuksen ongelmana on, ettei tallennusta voi rajata mikrofonilla pelkästään haluttuun puheeseen, muuten kuin asettamalla mikrofoni mahdollisimman lähelle puhujaa. Tämän vuoksi tunnistuksen haasteena on erottaa analysoitava puhe ympäristön muista äänistä, kuten liikenteen melusta, liikkumisesta syntyvistä äänistä ja etenkin taustalla kuuluvista muiden ihmisten puheesta. (Kurimo, M. 2009)

”Puheen eri äänteiden ominaispiirteiden laskemisessa 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 (foneemeja) parhaiten kuvaavat piirteet niin, että kaikki tunnistuksen kannalta ylimääräinen informaatio, kuten puhujan äänenkorkeus, painotukset ja ympäristön äänet, karsiutuu pois.” (Kurimo, M. 2009, 337) Jokaisesta ikkunoidusta signaalin pätkästä saatujen äänteiden ominaispiirteistä muodostetaan yksi piirrevektori, joka sisältää 39 ominaispiirrettä, jotka kuvaavat kyseisen signaalisegmentin spektrin sisällön, energian ja spektrin muutokset.

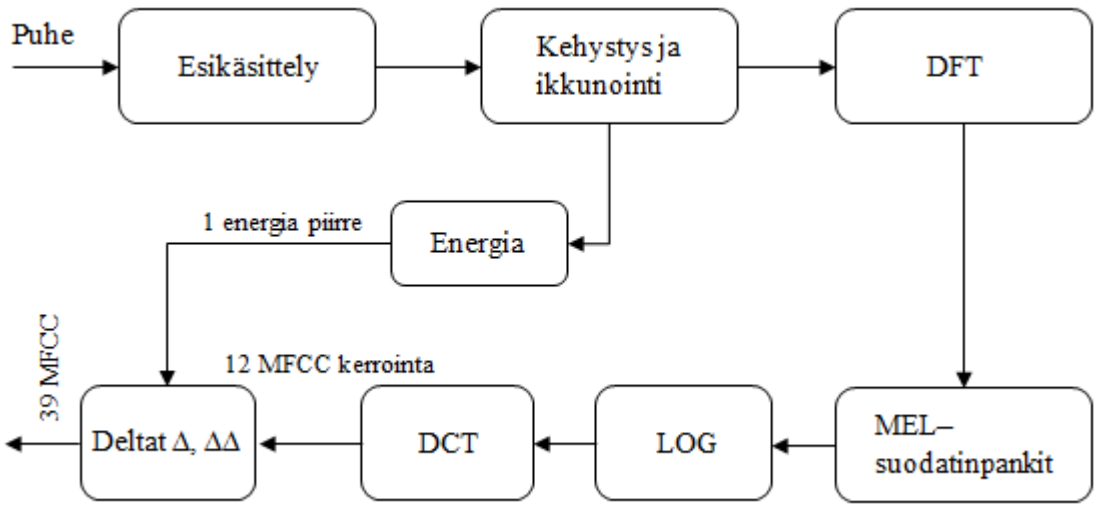
Akustisessa mallinnuksessa tai äänteiden tunnistamisen vaiheessa lasketaan todennäköisyydet, joilla puheesta irrotettu signaalisegmentti olisi peräisin tietystä foneemista. Jokaiselle HMM tilalle, joka vastaa foneemia tai sen rakenneosaa, lasketaan todennäköisyys, jolla tila tuottaa signaalia kuvaavan piirrevektorin. Yksinkertaistettu hahmotustapa tämän vaiheen ulostulolle on todennäköisyys vektorien sekvenssi, yksi jokaiselle aikaikkunalle, jokainen vektori kussakin aikaikkunassa sisältää todennäköisyydet joilla jokainen foneemi tai foneemin rakenneosa olisi tuottanut signaalia kuvaavan piirrevektorin kyseisellä hetkellä. Tästä saadut foneemitodennäköisyydet syötetään hakualgoritmille (tyypillisesti Viterbi–algoritmi), joka etsii kaikkein todennäköisimmän viestihypoteesin yhdistämällä foneemitodennäköisyydet sekä kielimallin antamien sanojen ja sanajonojen todennäköisyydet. (Jurafsky, D. & Martin, J.H. 2008) Tyypillisen ASR–järjestelmän yksinkertaistettu rakenne on esitetty kuviossa 1. Seuraavissa kappaleissa käsitellään yksityiskohtaisemmin tunnistusprosessin eri vaiheita.



KUVIO 1. ASR–järjestelmän rakenne. (Jurafsky, D. & Martin, J.H. 2008)

3.1 Puhesignaalin kuvaaminen ja mallintaminen piirrevektorina

Tämän osion tavoitteena on kuvata kuinka mitattu akustinen signaali muunnetaan piirrevektorisekvenssiksi laskemalla signaalista äänteitä kuvaavat ominaispiirteet, jossa kukin vektori esittää informaation signaalin lyhyestä aikaikkunasta. Piirteiden laskentaan on useita erilaisia hyväksi havaittuja tapoja, joista kaikkein yleisin vaihtoehto piirteiden ominaisuuksiksi on MEL–taajuuskepstrikertoimet (engl. Mel-Frequency Cepstral Coefficients, MFCCs), jotka lasketaan käyttämällä ikkunafunktiota (yleensä Hamming), Fourier- muunnosta, psykoakustisia suodatinpankkeja (MEL suodinpankit), logaritmista tiivistämistä ja diskreettiä kosinimuunnosta (DCT, kuvio 2).



KUVIO 2. 39–ulotteisen MFCC piirrevektorin muodostamisen lohkokaavio. (Gales, M. & Young, S. 2008; Kevin, M. 2008)

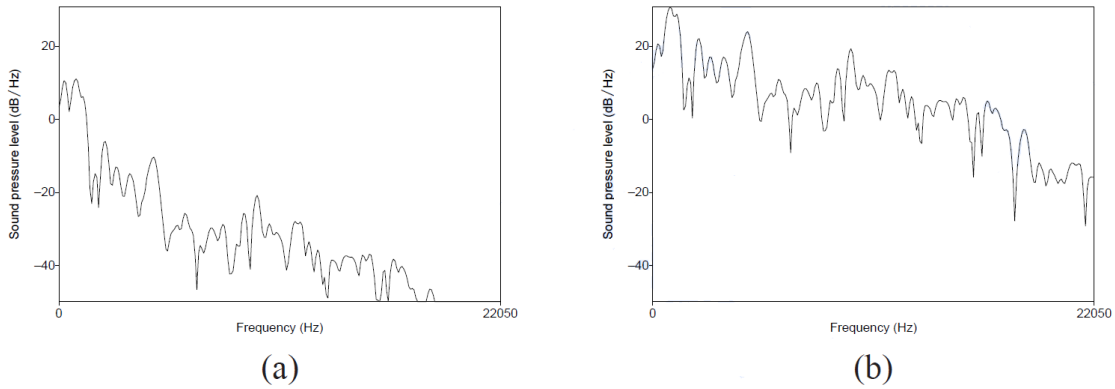
3.1.1 Puhesignaalin esikäsittely

Ensimmäinen vaihe puheen esikäsittelyssä on muuntaa analoginen puhesignaali digitaaliseen muotoon. Tämän analogia-digitaali-muunnoksen prosessissa on kaksi vaihetta: näytteenotto ja kvantisointi. Signaali näytteistetään ottamalla amplitudiarvo tietyn aikavälein; näytteenottotaajuus on otettujen näytteiden määrä sekunnissa. Jotta näytteistyksessä analoginen signaali saataisiin mitattua tarkasti, on syytä ottaa vähintään kaksi näytettä jaksoa kohti; mittaamalla aallon positiivinen osa ja aallon negatiivinen osa. Enemmän kuin kaksi näytettä per jakso lisää amplitudi tarkkuutta, mutta vähemmän kuin kaksi näytettä aiheuttaa laskostumista. Täten näytteenottoon vaadittava näytteenottotaajuus on oltava vähintään kaksi kertaa niin suuri kuin signaalin sisältämä suurin taajuuskomponentti (Nyquistin näytteenottoteoreema). Suurin osa ihmisen puheen sisältämä tieto on < 8000 Hz taajuusalueella, näin ollen 16000 Hz näytteenottotaajuus olisi tarpeellinen vaaditun tarkkuuden saavuttamiseksi. (Jurafsky, D. & Martin, J.H. 2008)

Puhesignaalin näytteistyksen ja kvantisoinnin jälkeen tehostetaan korkeataajuisten komponenttien tehoa suodattamalla signaali. Tarkastelemalla äänteen /aa/ spektri otetta (kuvio 3) huomataan kuinka äänteen alemmilla taajuuksilla on enemmän energiaa kuin korkeammilla taajuuksilla. Tehostamalla korkeamman taajuusalueen tehoa, saadaan näiden korkeampien formanttien informaatio paremmin saataville akustiseen mallinnukseen, joka parantaa foneemien havaitsemisen tarkkuutta. Digitalisoidun puhesignaalin suodatus tehdään ensimmäisen kertaluvun ylipäästösuodattimella:

$$y[n] = x[n] - ax[n-1], \tag{1}$$

jossa $x[n]$ on diskreettiaikaisen digitaalisen signaalin näyte ja $0.9 \leq a \leq 1.0$. Kuviossa (3) on esitettynä esimerkki äänteen /aa/ spektristä ennen ja jälkeen suodatusta. (Jurafsky, D. & Martin, J.H. 2008; Ursin, M. 2002)

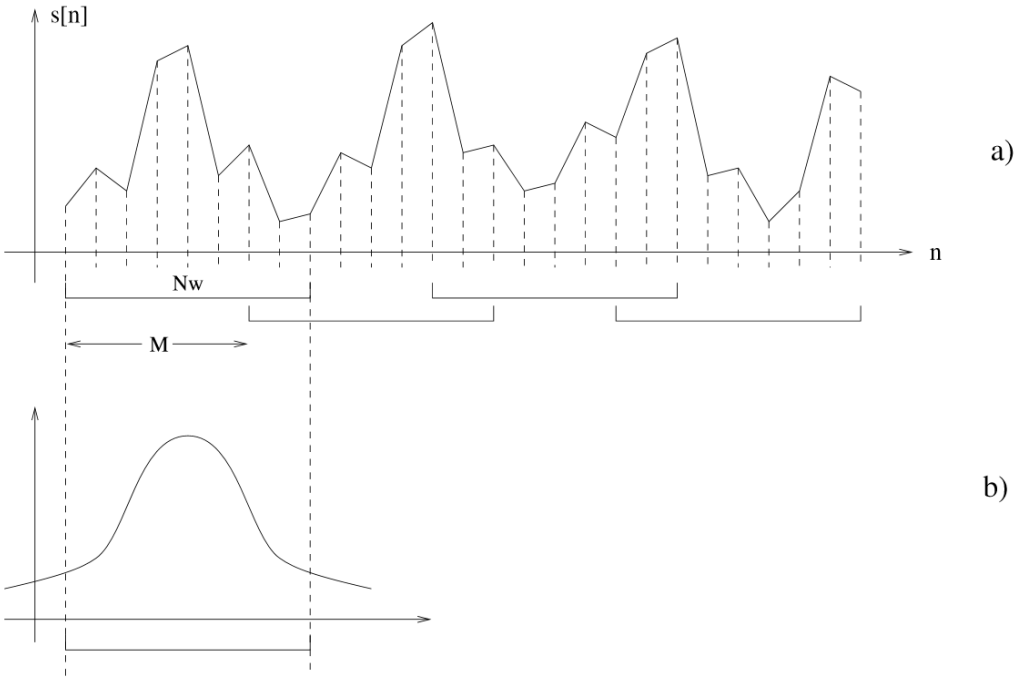


KUVIO 3. Ote äänteen /aa/ spektristä ennen (a) ja jälkeen (b) suodatusta. (Jurafsky, D. & Martin, J.H. 2008)

3.1.2 Ikkunointi ja Fourier–muunnos

Ensimmäinen vaihe MFCC piirrevektorin muodostamisessa on mikrofonilla tallennetun ja digitalisoidun signaalin jako lyhyisiin vain kymmenkunta millisekunnin (yleensä 25 ms) kehyksiin, jotka sisältävät N määrän näytteitä. Jokaista näistä puheesta otettua kehystä käsitellään itsenäisinä signaaleinaan. Jotta vältyttäisiin olennaisten tietojen puuttumiselta, peräkkäiset kehykset on sijoitettu osittain päällekkäin siten, että ensimmäinen näyte kussakin kehyksessä on tyypillisesti asetettu 10 millisekunnin välein, M näyte määrän verran (kuvio 4). (Gales, M. & Young, S. 2008; Kevin, M. 2008)

Puhesignaalin paloittelusta saaduille jokaiselle yksittäiselle kehykselle suoritetaan ikkunafunktio, jolla tasoitetaan kehyksen reunat, jotka muutoin aiheuttaisivat suurtaajuisia komponentteja esiintymään spektrissä. Tästä syystä yleisemmin käytetty ikkunafunktio MFCC piirreirrotuksessa on Hamming ikkunafunktio, joka vaimentaa kehyksen signaalin alun ja lopun amplitudiarvot lähelle nollaa, jotta spektriin ei tule säröä epäjatkuvuudesta johtuen. (Jurafsky, D. & Martin, J.H. 2008)



KUVIO 4. Piirrevektorin muodostamisen kaksi ensimmäistä vaihetta. a) näytteistetyn signaalin jako kehyksiin. b) jokaisen yksittäisen kehyksen ikkunointi Hamming–funktiolla. (Giampiero, S)

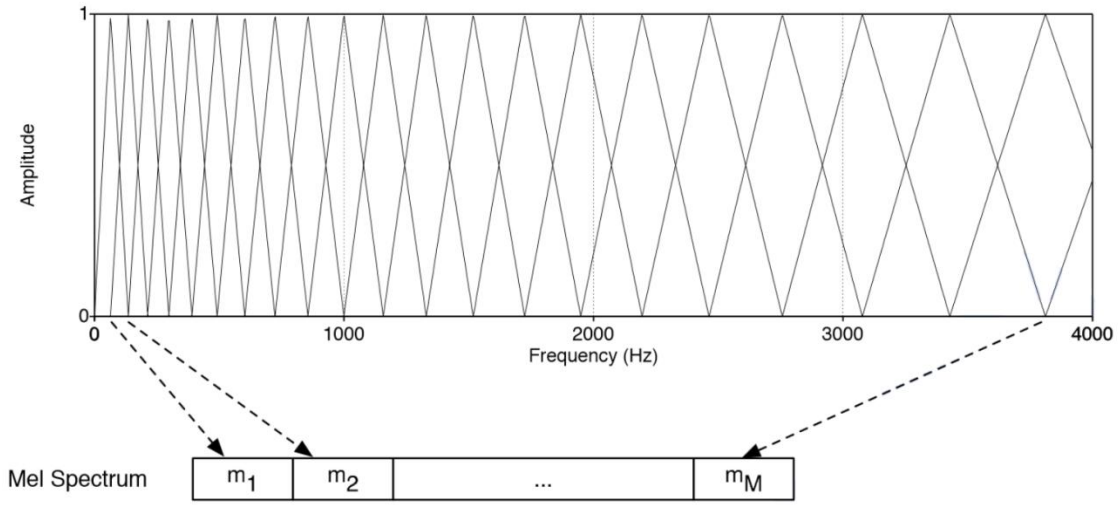
Seuraavassa käsittelyvaiheessa poimitaan spektraalinen informaatio ikkunoidusta signaalista, josta saadaan selville kuinka paljon energiaa signaali sisältää eri taajuusalueilla. Jokaisesta lyhyestä kehyksestä eli aikaikkunasta lasketaan taajuusjakauma käyttäen Diskreettiä Fourier-muunnosta (DFT), joka muuntaa kunkin kehyksen esityksen aikatasosta taajuustasolle. Diskereettiaikaisen jaksollisen signaalin $x[n]$ (jakso N) Fourier–muunnos määritellään kaavalla:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-i2\frac{\pi}{N}kn}, \qquad k = 0, \dots, N-1 \tag{2}$$

Yleisesti käytetty algoritmi DFT:n laskemiseen on FFT (Fast Fourier Transform) eli nopea Fourier'n muunnos, joka on laskennallisesti nopeampi toteutus. (Jurafsky, D. & Martin, J.H. 2008)

3.1.3 MEL–suodatinpankit ja logaritmisointi

Taajuustason ikkunoille käytetään matemaattisia operaatioita vähentämään signaalin tarpeetonta dataa. Signaalin dataa voidaan vähentää merkittävästi käyttämällä MEL–asteikon suodatinpankkeja (kuvio 5), jotka pyrkivät mallintamaan ihmiskorvan taajuusherkkyyttä. Tavoitteena on poimia jokaisesta ikkunasta puheen äänteitä kuvaavat piirteet siten, että tunnistuksen kannalta kaikki ylimääräinen tieto, kuten ympäristön äänet ja puhujan äänenkorkeus karsiutuu pois. Ihmisen kuulo ei ole yhtä herkkä kaikilla taajuusalueilla, vaan se on vähemmän herkkä korkeammilla taajuuksilla, suunnilleen yli 1000 Hz taajuuksilla. Mallintamalla tätä ihmisen kuulo ominaisuutta piirreirrotuksessa, saadaan parannettua puheen tunnistusta. MEL–asteikon ensimmäiset 10 suodatinta on asetettu erilleen toisistaan lineaarisesti 1000 Hz alapuolelle, ja loput suodattimet levitetty logaritmisesti 1000 Hz yläpuolelle. (Jurafsky, D. & Martin, J.H. 2008)



KUVIO 5. Ihmisen kuulojärjestelmää simuloivien kolmiosuodatinten sijoitus taajuusasteikolle MEL–asteikon mukaisesti. Jokainen kolmiosuodin kerää energian annetulta taajuusalueelta. (Jurafsky, D. & Martin, J.H. 2008)

MEL–suodatinpankki on yksinkertaisesti joukko limittäisiä kolmiovasteisia kaistanpäästö suodattimia taajuustasolla. Ensimmäinen suodatin on hyvin kapea ja antaa viitteitä siitä, kuinka paljon energiaa esiintyy 0 Hz:n lähettyvillä. Taajuuden kasvaessa suodattimien vasteet kasvavat. MEL–suodatinpankin jokainen kolmiosuodin kerää energian annetulta taajuusalueelta. Suodinpankin energioiden laskemisessa jokaista kaistaa kohti lasketaan yksi arvo, joka saadaan painotettuna keskiarvona kaistan sisältämistä energioista. Painofunktiona käytetään edellä kuvattuja kolmio suodattimia (käytetään yleensä

noin 20:ntä suodinta). Suodinpankin ulostulona saadaan 20 numeroarvoa/kehys, joista yleensä taltioidaan vain ensimmäiset 12 kepstri arvoa. Tästä saadut MEL–tehospektrikertoimet m_k logaritmisoidaan ($\log m_k$) kertoimiksi S_k , koska yleisesti ottaen ihmisen kuuloherkkyys signaalintasoon on logaritminen; ihmiset ovat vähemmän herkempiä pieniin amplitudieroihin suurilla amplitudeilla kuin matalilla amplitudeilla. Tämän lisäksi logaritmisointi tekee piirteiden arvioinnin vähemmän alttiimmaksi puhujan äänen tason vaihteluille, joka voi johtua esimerkiksi puhujan liikkumisesta lähemmäksi tai kauemmaksi mikrofonista. (Gales, M. & Young, S. 2008; Jurafsky, D. & Martin, J.H. 2008)

3.1.4 Diskreetti kosinimuunnos

Vaikka pelkästään logaritmisia MEL–spektrikertoimia olisi mahdollista käyttää itsenään piirteiden esityksenä foneemien tunnistuksessa, on kepstrikertoimilla useita hyödyllisiä prosessointi etuja ja ne myös parantavat huomattavasti foneemien tunnistusta. MEL–kepstrikertoimet (MFCCs) voidaan määrittää laskemalla diskreetti kosinimuunnos (DCT) suodatinten ulostulojen logaritmeista S_k, käyttäen seuraavaa yhtälöä:

$$C_n = \sum_{k=1}^{K} S_k \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right], \quad n = 0, 1, 2, \dots, N, \tag{3}$$

jossa n on kepstraalikertoimen indeksiarvo, N on haluttu kepstraalisten kerrointen lukumäärä, jonka arvo on tyypillisesti 12 ja S_k, $k = 1, 2, \dots, K$ on K–kanavaisen suodatinpankin logaritmoitu ulostulo indeksillä k. Lopputuloksena tästä on 12 kepstrikerrointa kutakin kehystä kohti. (Gales, M. & Young, S. 2008; Jurafsky, D. & Martin, J.H. 2008)

3.1.5 Kehyksen energia ja deltat

Kepstrikertoimien erotus edellisessä osiossa diskreetin kosinimuunnoksen avulla tuottaa 12 kepstrikerrointa jokaista kehystä kohden. Kepstrikertoimien laskemisen jälkeen, lisätään kehyksen MFCC vektoriin kolmastoista piirre: kehyksen energia. Kehyksen energia korreloi foneemin identiteetin kanssa, ja on siten hyödyllinen indikaattori foneemin tunnistamiseen. Kehyksen energia on kehyksen näytteiden tehojen summa:

$$E(f) = \sum_{t=t_1}^{t_k} x_f^2\,[t], \tag{4}$$

jossa $x_f[t]$ on kehystetyn signaalin t:nnen näytteen arvo kehyksessä f, ja t_1 on kehyksen ensimmäinen näyte ja t_k on näytteiden lukumäärä kehyksessä. (Jurafsky, D. & Martin, J.H. 2008)

Toinen tärkeä seikka puhesignaalissa on, että se ei ole vakio kehyksestä kehykseen. Tämän vuoksi on tarvetta lisätä vektoriin myös piirteitä, jotka kuvaavat kepstraalisten piirteiden muutosta kehysten välillä. Näitä piirteitä kutsutaan delta (nopeus piirre) ja delta–delta (kiihtyvyys piirre) kertoimiksi. Jokaista kehyksen 13:a piirrettä kohden (12 kepstrikerrointa + kehyksen energia) lisätään delta ja delta–delta piirre. Jokainen 13:sta delta piirteestä edustaa kehysten välistä muutosta vastaavissa kepstri / energia piirteissä, kun taas jokainen 13:sta delta–delta piirteestä edustaa kehysten välistä muutosta vastaavissa delta piirteissä. Nämä ensimmäisen ja toisen kertaluvun delta-kertoimet kuvaavat puhesignaalin dynaamisia ominaisuuksia, jotka ovat olennaisia mallintamaan foneemin siirtymistä toiseen. Ensimmäisen kertaluvun deltat saadaan laskettua kehyksen kepstrikertoimista C seuraavasti:

$$\Delta C_t = \frac{\sum_{i=1}^{J} i(C_{t+i} - C_{t-i})}{2\sum_{i=1}^{J} i^2}, \tag{5}$$

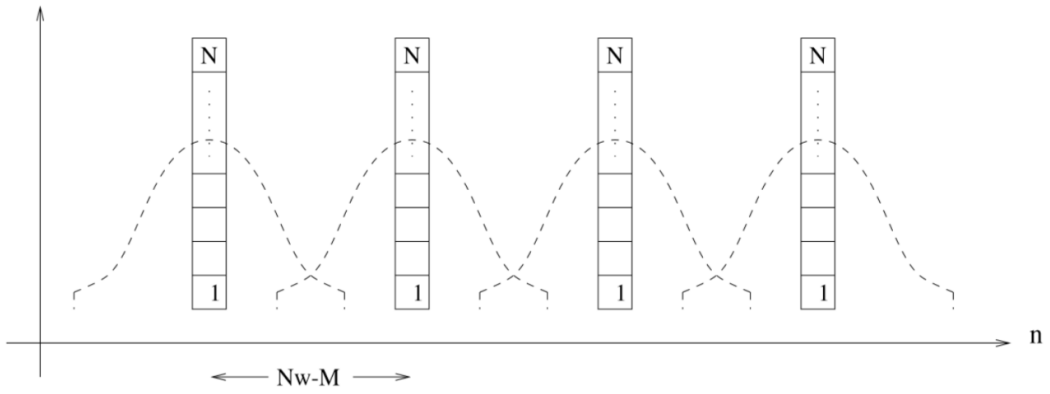
jossa ΔC_t on kepstrikertoimesta C_t laskettu delta kerroin hetkellä t ja parametri J on ikkunan leveys jolle delta-kertoimet lasketaan, tyypillisesti J =2. Toisen kertaluvun (delta–delta) kertoimet $\Delta^2 C_t$ lasketaan ensimmäisen kertaluvun deltoista käyttäen yhtälöä (5) samassa muodossa, mutta korvaamalla kepstrit C_t deltoilla ΔC_t. Kehyksen energian delta ja delta–delta kertoimet voidaan määrittää samalla menetelmällä kuin kepstrien deltat. (Gales, M. & Young, S. 2008; Jurafsky, D. & Martin, J.H. 2008; Ursin, M. 2002)

3.1.6 Piirrevektorin rakenne

Kehyksen energian ja sitten delta ja delta–delta piirteiden lisääminen 12:een kepstri piirteeseen tuottaa kutakin kehystä kohden 39 MFCC piirrettä:

- 12 kepstrikerrointa
- 12 delta kepstrikerrointa
- 12 delta–delta kepstrikerrointa
- 1 energia kerroin
- 1 delta energia kerroin
- 1 delta–delta energia kerroin

Näitä piirteitä käytetään akustisessa mallinnuksessa, joka on kuvattuna osiossa 3.2, määrittämään puhesignaalissa esiintyvien foneemien todennäköisyydet. Puheen piirteiden mallintamisen lopputuloksena muodostetaan jokaisesta analysoidusta puheen lyhyestä osasta eli ikkunoidusta signaalin kehyksestä yksi piirrevektori (kuvio 6). Piirrevektori sisältää edellä esitetyt 39 lukuarvoa, jotka on valittu kuvaamaan kehyksessä esiintyvä puheentunnistuksen kannalta merkittävä sisältö mahdollisimman kompaktissa muodossa. (Jurafsky, D. & Martin, J.H. 2008; Kurimo, M. 2009)



KUVIO 6. Ikkunoitujen signaalikehysten piirrevektorit. (Giampiero, S)

3.2 Äänteiden tilastollinen mallintaminen

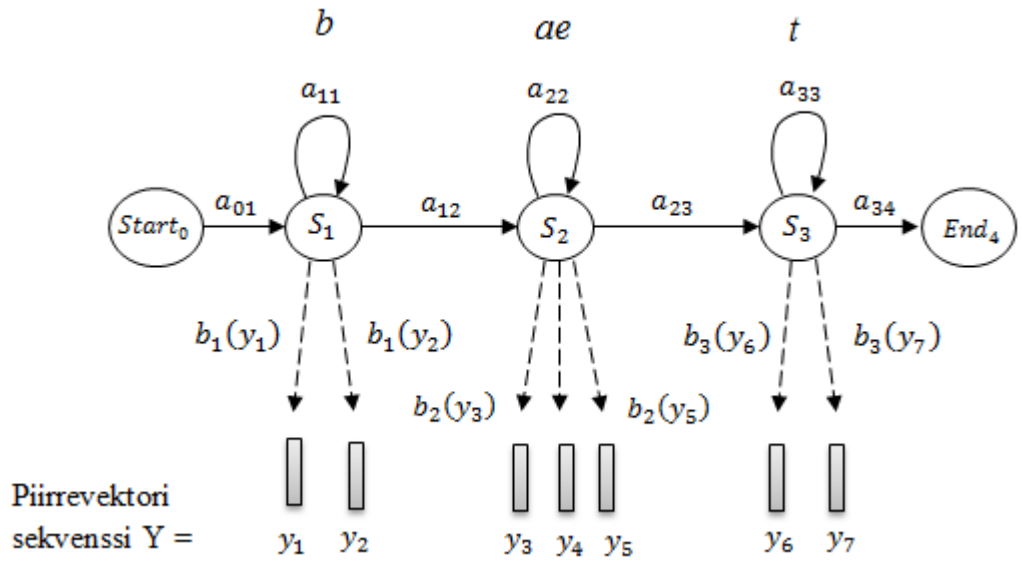
Puheen akustinen mallintaminen viittaa prosessiin, jonka avulla voidaan laskea todennäköisyydet, joilla puhesignaalista erotettu piirrevektori olisi peräisin tietystä foneemis-

ta. Akustisessa mallinnuksessa käytetään erilaisia tilastollisia hahmontunnistustekniikoita apuvälineinä, joista nykypäivänä käytetyin tilastollinen mallinnus perustuu Markovin piilomalliin (engl. HMM eli hidden Markov model).

Akustinen malli sisältää tilastolliset esitykset jokaisesta erillisestä äänteestä eli foneemista, jotka muodostavat tunnistettavan kielen sanaston. Esimerkiksi englanninkielessä on noin 40 erillistä äännettä, jotka ovat hyödyllisiä puheentunnistuksessa. Akustinen malli luodaan ottamalla suuri tietokanta puhetta (puhekorpus) ja käyttämällä tiettyjä koulutus algoritmeja muodostamaan tilastolliset mallit jokaiselle kielen foneemille. Näitä tilastollisia malleja kutsutaan Markovin piilomalleiksi (HMM), jotka ovat todennäköisyyteen perustuvia tilakoneita, ja joiden koulutetut parametrit koostuvat todennäköisyysjakaumista $b_j()$ ja siirtymätodennäköisyyksistä a_{ij}. Markovin piilomallin nimi kuvaa sitä, että tilan vaihtuminen ei ole suoraan havaittavissa vaan sen on niin sanotusti kätketty ja nämä piilossa olevat siirtymätodennäköisyydet pyritään päättelemään tiloista havaittujen lopputulemien perusteella. Myös tilan ominaisuudet eli jakaumamalli ja kesto oletetaan riippumattomiksi edellisistä ja seuraavista tiloista. ”Äänteiden tilastollisten mallien muodostamisessa kullekin äänteelle määritetty todennäköisyysjakauma kuvaa piirrevektorien esiintymistä äännettä vastaavassa tallenteen osassa. Tavallisesti jakauma mallinnetaan moniulotteisella normaalijakaumalla (GMM), jossa jokaiselle piirrevektorin alkiolle on suuren puheaineiston perusteella estimoitu keskiarvo ja keskihajonta.” (Kurimo, M. 2009) Todennäköisyysjakauman avulla pystytään Mel–kepstri-piirteistä laskemaan todennäköisyys, jolla puheesta erotettu signaalikehys olisi peräisin tietystä foneemista. Jokaisella foneemilla on oma HMM. (Gales, M. & Young, S. 2008; Kurimo, M. 2009)

Markovin piilomalleja (HMMs) voidaan käyttää puheen mallintamisessa usein eri tavoin. Hyvin yksinkertaisiin tunnistus järjestelmiin, kuten numeroiden tunnistaminen tai kyllä-ei sanojen tunnistukseen, voidaan rakentaa HMM jonka tilat vastaavat kokonaisia sanoja. Suuremmissa tunnistus järjestelmissä HMM tilat vastaavat foneemi yksiköitä ja sanat muodostuvat näiden foneemien sekvensseistä. Kuviossa (7) on kuvattuna tyypillinen puheentunnistuksessa käytetty vasemmalta oikealle (left-to-right) HMM ketju sanalle ”bat”, jossa erikseen mallinnetaan piirteiden tiheysfunktiot systeemin eri tiloissa ja tilojen välisten siirtymien todennäköisyydet. Siirtymä todennäköisyys $a_{ij} = P(S_t = j \mid S_{t-1} = i)$, jossa S_t on tila indeksi ajalla t, on todennäköisyys siirtyä tilasta i tilaan j huomioiden edellisen tilan i, tai siirtyä takaisin samaan tilaan (self-loop), joka

mahdollistaa yksittäisen foneemin toiston. Silmukoiden avulla voidaan mallintaa foneemien vaihtelevia kestoja; pidemmät äänteet vaativat enemmän luuppeja. Kaikille mahdollisille tila siirtymille a_{ij} on estimoitu oma siirtymä todennäköisyytensä, jotka ovat määritetty ns. siirtymätodennäköisyys matriisissa $A = [a_{01}a_{02}a_{03} \dots a_{n1} \dots a_{nn}]$. Nämä siirtymä todennäköisyydet sekä todennäköisyysjakaumat saadaan hyvin estimoitua puhekorpuksen äänteiden koulutus aineistosta, joka koostuu tunnistimelle annettujen lausahduksien äänisignaaleista ja niiden oikein tulkinnoista. Tilojen väliset siirtymätodennäköisyydet ja HMM tilat muodostavat yhdessä ääntämissanakirjan; HMM tilakaavio rakenne jokaiselle sanalle, jonka tunnistin kykenee tunnistamaan. (Deng, L. & Huang, X. 2009; Jurafsky, D. & Martin, J.H. 2008)



KUVIO 7. Yksinkertainen foneemi-tilainen HMM ketju sanalle "bat", jossa jokainen HMM tila vastaa yksittäistä foneemia. (Gales, M. & Young, S. 2008)

Vertaamalla eri äänteiden tilamalleja uudesta puhenäytekehyksestä laskettujen piirrevektoreiden sekvenssiin, voidaan jokaista tilaa kohti laskea todennäköisyysjakauma, joka määrittää miten todennäköisesti malli voisi generoida tämän näytteen. Todennäköisyys, jolla tila j tuottaa signaalia kuvaavan piirrevektorin y, saadaan yleensä usean muuttujan Gaussin mikstuurimallista (GMM), joka on usean muuttujan normaalijakauman painotettu summa:

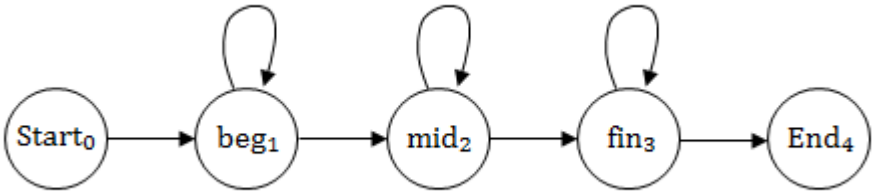
$$b_j(y) = p(y \mid S_j) = \sum_{m=1}^{M} c_{jm} N(y; \mu_{jm}, \Sigma_{jm}), \quad (6)$$

jossa mikstuurin painot täyttävät ehdot: $c_{jm} \geq 0$ ja $\sum_{m=1}^{M} c_{jm} = 1$. Gaussin moniulotteinen normaalijakauma $N(y; \mu_{jm}, \Sigma_{jm})$ määritetään kaavalla:

$$N(y; \mu_{jm}, \Sigma_{jm}) = \frac{1}{\sqrt{(2\pi)^D |\Sigma_{jm}|}} e^{-\frac{1}{2}(y-\mu_{jm})^T \Sigma_{jm}^{-1}(y-\mu_{jm})}, \quad (7)$$

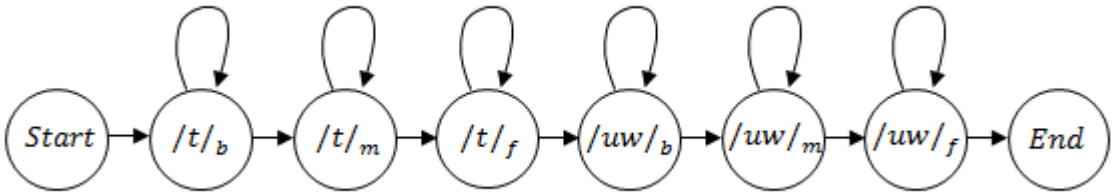
jossa $\mu_{jm,}$on m:n gaussisen komponentin keskimääräinen vektori dimensiolla D ja Σ_{jm} on kovarianssimatriisi. Kuten 3.1 osiossa käytiin läpi, tyypillinen MFCC piirrevektorin dimensio (D) LVCSR järjestelmässä on 39. (Deng, L. & Huang, X. 2009; Kurimo, M. 2008)

Yksinkertaisiin puheentunnistus tehtäviin, yksittäisen HMM tilan käyttö foneemin esittämiseen on riittävä. Yleisesti laajan sanaston jatkuvan puheentunnistuksen tehtäviin tarvitaan kuitenkin hienojakoisempi foneemimalli, koska yksittäisen mallinnettavan foneemin vastaavan tallennetun puhesignaalin osan piirteet ovat usein erilaisia foneemin alussa, keskivaiheilla ja lopussa, joka on otettava huomioon foneemeille rakennettavassa piirteiden tilastollisessa mallissa. Yksittäiset foneemit voivat kestää jopa yli yhden sekunnin eli yli 100 kehystä, mutta nämä kehykset eivät ole akustisesti identtisiä. Foneemin spektriominaisuudet ja energia määrä vaihtelevat huomattavasti foneemin eri osissa. Tämän vuoksi foneemimalli rakennetaan useasta peräkkäisestä tilasta (yleensä kolmesta), joilla jokaisella on oma jakauma- ja kestomallinsa. Yleisesti laajan sanavaraston jatkuvan puheentunnistusjärjestelmän (LVCSR) yksittäisen foneemin rakenteena käytetään kolmea HMM tilaa: alkuosa, keskiosa ja loppuosa (jotka vastaavat foneemiin siirtymistä, vakaata tilaa ja pois siirtymistä). Jokainen foneemi täten koostuu kolmesta emittoivasta HMM tilasta (plus kahdesta ei-emittoivasta tilasta kummassakin päässä, start ja end) yhden sijaan (kuvio 8). (Jurafsky, D. & Martin, J.H. 2008)



KUVIO 8. Tyypillinen viiden tilan HMM malli foneemille, joka koostuu kolmesta emittoivasta tilasta ja kahdesta ei-emittoivasta tilasta. (Jurafsky, D. & Martin, J.H. 2008)

Tunnistuksen aikana jokaiselle annetulle sanalle w_k, vastaava HMM rakenne syntetisoidaan ketjuttamalla foneemimalleja yhteen muodostamaan kokonaisia sanoja, jossa kunkin sanan w_k HMM rakenne saadaan leksikossa olevien sanojen ääntämismallista. Käyttäen edellä kuvattua 3–tilan foneemi mallia, jossa yksittäisen foneemin rakenteena käytetään kolmea HMM tilaa: alkuosa, keskiosa ja loppuosa, voidaan kokonaisen sanan HMM ketjun rakentaminen yksinkertaisesti toteuttaa korvaamalla ei-emittoivat start ja end tilat foneemimallilla josta on suora yhteys edellisen ja seuraavan äänteen emittoiviin tiloihin, jättäen vain kaksi ei-emittoivaa tilaa koko sanalle. Jokaiselle sanalle muodostuva HMM rakenne on yksinkertaisesti foneemimallien ketju, jossa jokainen foneemi koostuu kolmesta tilasta (Kuvio 9). (Gales, M. & Young, S. 2008; Jurafsky, D. & Martin, J.H. 2008)

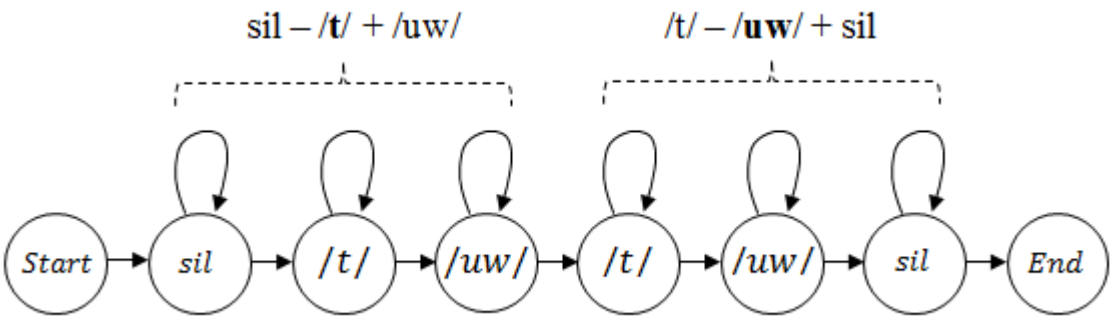


KUVIO 9. Yhdistetty ääntämismalli sanalle ”two” [t uw], joka on muodostettu yhdistämällä kaksi foneemimallia, joissa molemmissa on kolme emittoivaa tilaa. (Jurafsky, D. & Martin, J.H. 2008)

3.2.1 Kontekstisidonnaiset akustiset mallit

Ongelmana edellä kuvatuissa kontekstista riippumattomissa akustisissa malleissa (jossa yksittäisen foneemin rakenteena käytetään kolmea emittoivaa HMM tilaa) on, että fo-

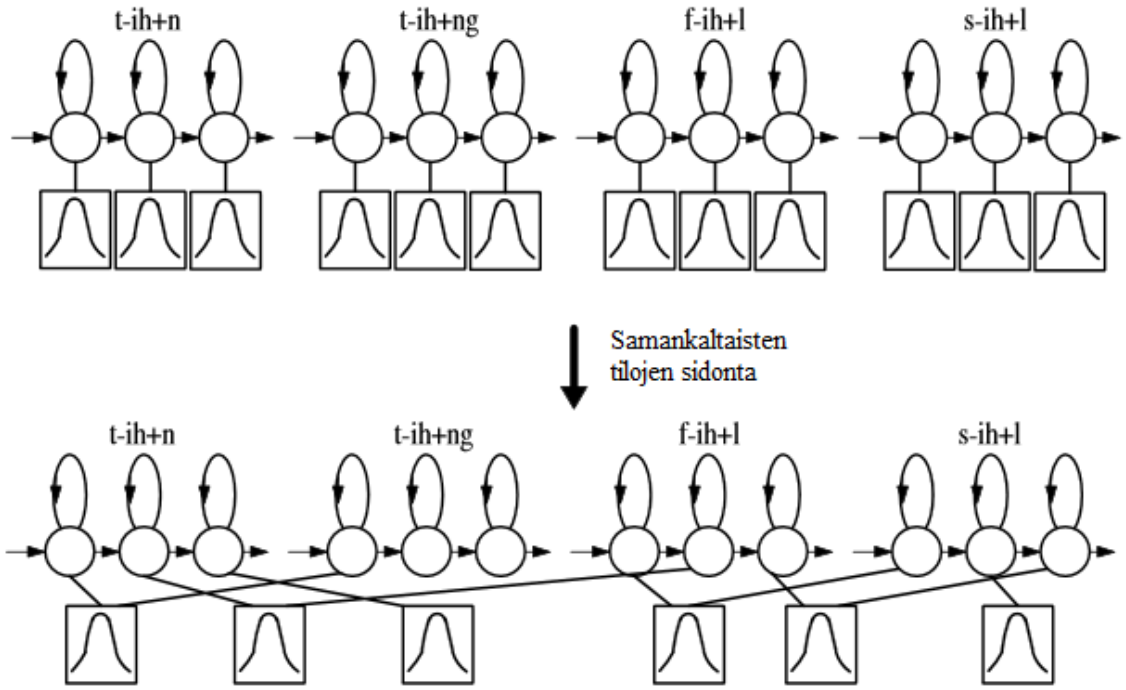
neemien akustiset piirteet vaihtelevat edellisen ja seuraavan foneemin perusteella. Jotta voidaan mallintaa vaihtelu, joita foneemissa esiintyy eri konteksteissa, useimmat LVCSR järjestelmät korvaavat konteksti riippumattomat (CI) äännemallit kontekstisidonnaisilla (CD) foneemimalleilla. Kaikista yleisin kontekstisidonnainen malli on trifoni HMM, joka esittää foneemin sen vasemmassa ja oikeassa kontekstissa. Esimerkiksi trifoni [b – ae + t] tarkoittaa foneemin [ae] yhteydessä puhuttua edellistä foneemia [b] ja seuraavaa foneemia [t]. Nimestään huolimatta, trifoni on yksinkertaisesti yksittäisen foneemin malli joka on esitettynä sen lähinaapurien kontekstissa. Tilanteissa joissa trifonilla ei ole täyttä kontekstia, käytetään bifoni mallia, joka mallintaa foneemin sen vasemmassa (edellisessä) tai oikeassa (seuraavassa) kontekstissa. Esimerkiksi bifoni [a–b] tarkoittaa että foneemia [b] edeltää [a] ja [b+c] tarkoittaa että foneemia [b] seuraa [c]. Kontekstisidonnaiset foneemit kaappaavat tärkeän osan foneemien variaatiosta ja ne ovat olennainen osa modernia ASR järjestelmää. (Jurafsky, D. & Martin, J.H. 2008; Ursin, M. 2002)



KUVIO 10. Kaksi yhdistettyä trifoni HMM mallia sanalle ”two” [t uw]. ’Sil’ tarkoittaa hiljaisuutta sanan alussa ja lopussa, joka on myös mallinnettu ’foneemina’.

Äännemallinnuksessa jokaiselle trifonille opetetaan kolmitilainen Markovin piilomalli–ketju emissiotodennäköisyysjakaumineen ja siirtymätodennäköisyyksineen. Trifoneihin perustuvan akustisen mallinnuksen ongelmaksi muodostuu puolestaan riittämätön opetusaineisto, sillä puhutunkielen koostuessa N määrästä foneemeja, on loogisesti puolestaan olemassa N^3 potentiaalista trifonia. Näin ollen on epätodennäköistä, että monelle trifoni mallille olisi riittävää koulutusmateriaalia luotettavien parametrien estimointiin. Tämän lisäksi hyvin suuren trifoni joukon koulutus johtaisi hyvin monimutkaiseen tunnistimeen, ja tunnistusprosessin hidastumiseen. Käytännössä jokainen kolmen foneemin sekvenssi ei kuitenkaan ole mahdollinen tai ne ovat hyvin harvinaisia, ja koartikulaatiosta huolimatta jotkin trifonit ovat melko samankaltaisia, jolloin ne on parempi mallin-

taa samalla mallilla. Yleisin ratkaisu koulutettavien trifoni parametrien vähentämiseen on jakamalla joidenkin mallien parametrit sitomalla tilojen todennäköisyysjakaumat muiden samankaltaisten tilojen kanssa. Sitomalla kaksi tilaa toisiinsa tarkoittaa, että ne jakavat saman jakauman. Esimerkki trifonien tilojen klusteroinnista on esitettynä kuviossa 11, jossa Gaussin jakaumat on jaettu useiden eri trifoni HMM tilojen kesken. (Jurafsky, D. & Martin, J.H. 2008; Ursin, M. 2002)



KUVIO 11. Esimerkki foneemin /ih/ eri trifoni tilojen klusteroinnista. (Gales & Young, 2008, s. 207)

3.3 Leksikko ja kielen mallintaminen

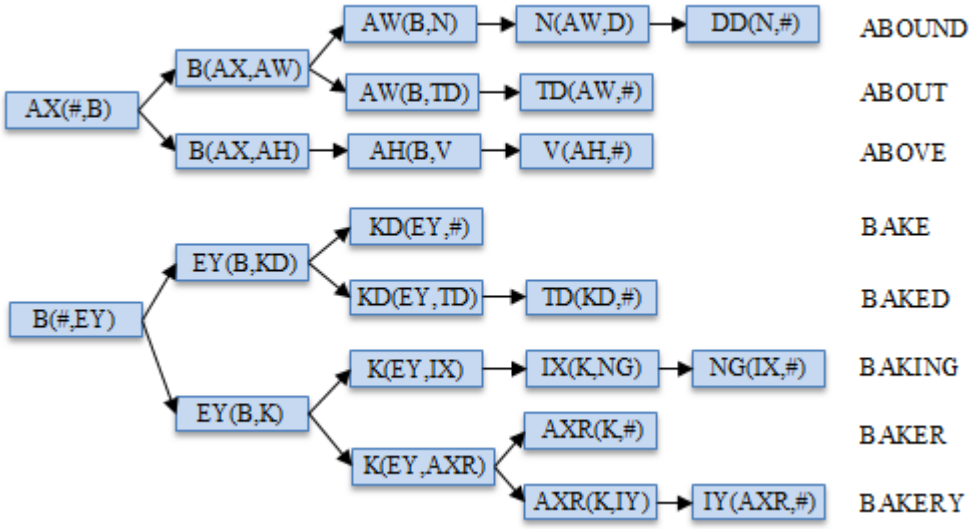
Yleisesti ottaen puheentunnistusjärjestelmän akustinen vaihe tuottaa joukon foneettisia todennäköisyyksiä, joita tunnistuksen aikana sovitetaan leksikossa olevien sanojen ääntämismalleihin, muodostaen optimaalisimman tilajonon. Tämän vaiheen aikana on tarpeellista ottaa käyttöön sääntöjä, jotka voivat kuvata kielellisiä rajoituksia joita luonnollisessa kielessä esiintyy ja joilla voidaan ratkaisevasti rajoittaa läpikäytävien vaihtoehtojen määrää sekä erotella toisistaan samalta kuulostavat sanat (homonyymit).

Teoriassa pelkkien äännemallienkin avulla voitaisiin tunnistaa puhetta muuttamalla puhesignaalista saadut äänteet sanoiksi. Tämä on kuitenkin käytännössä osoittautunut virhe herkäksi toteutustavaksi, koska kielessä on usein sanoja, joiden äännejono on lähes tai täsmälleen sama, vaikka niiden kirjoitusasu onkin erilainen. Esimerkkinä tästä on englanninkielessä sanat ”I” ja ”eye”, joiden äännejono on täsmälleen sama. Tällaisten sanojen tunnistus on mahdollista vain viereisten sanojen eli kontekstin perusteella.

Kehittyneimmille puheentunnistusjärjestelmille onkin opetettu sanasto ja kielen tilastollinen rakenne, jotta ne tietävät minkälaiset lauseet ovat järkeviä. Esimerkiksi, jos käyttäjä sanoo "thank" ja seuraavaksi sanan, joka kuulostaa sanalta "dew", niin tunnistin voi tehdä tilastollisen johtopäätöksen, että puhuja tarkoitti todennäköisimmin sanaa ”you”. Tätä prosessia, joka antaa todennäköisyydet sanoille ja sanayhdistelmille, kutsutaan puheentunnistuksessa kielimalliksi.

Tilastollisen kielimallin rakennus aloitetaan laatimalla sanasto eli leksikko ja määrittämällä jokaisen sanan esiintymistodennäköisyys ja todennäköisin ääntämismalli edellisessä osiossa kuvattujen tilamallien (foneemimallien) jonona. Joillakin sanoilla voi myös olla ääntämistavan useita eri variaatioita, jolloin niiden yhteyteen on myös määritettävä kunkin ääntämistavan esiintymistodennäköisyys. Koska leksikon koko lisää kuitenkin virheellisten tunnistusten määrää, on parasta karsia pois kaikista harvinaisimmat ääntämistavat, etenkin jos ääntämistavat ovat lähes samanlaisia. Useissa kielissä, kuten suomenkielessä, puheentunnistusjärjestelmän sanasto voi kuitenkin kasvaa erittäin suureksi, koska on huomioitava myös sanojen kaikki mahdolliset taivutusmuodot. Ääntämissanakirjan eli leksikon avulla voidaan määrittää mitä sanoja on olemassa ja mitkä foneemi yhdistelmät antavat tunnistuksessa kelvollisia sanoja. (Kurimo, M. 2009)

Leksikon sisältämien foneemi yhdistelmien organisointi toteutetaan laajan sanaston puheentunnistuksessa yleensä usean erillisen ääntämismallin sijasta foneemi verkostolla, jossa verkoston eri polut ilmaisevat tunnistettavia sanoja. Useat reaaliaikaisten tunnistimien nopeista hakualgoritmeista perustuvat puutietorakenteisen leksikon käyttöön, jossa sanojen ääntämistavat on organisoitu siten, että foneemit voidaan jakaa samankaltaisilla foneemi sekvenssillä alkavien sanojen kesken. Kuviossa 12 on esitettynä esimerkkiote puurakenteisesta leksikosta, jossa jokainen lehti/polku vastaa leksikossa olevaa sanaa. (Jurafsky, D. & Martin, J.H. 2008)



KUVIO 12. Esimerkki puutietorakenteisesta leksikosta, jossa jokainen solmu edustaa kolmetilaista trifonia ja foneemi verkoston eri polut edustavat sanakirjan eri sanoja. (Jurafsky, D. & Martin, J.H. 2008)

Kielimallin sisältävien sanayhdistelmien todennäköisyydet perustuvat yleensä suureen tekstiaineistoon eli korpukseen (koostuu yleensä miljoonista sanoista), jonka aineisto voi olla peräisin esimerkiksi sanomalehdistä, TV–ohjelmien tekstityksistä, kirjoista, Wikipedia artikkeleista, jne. Kielimallin tehtävänä on ennustaa sanojen esiintymistodennäköisyys toinen toisensa jälkeen tietyllä kielellä, kun N edellistä sanaa tunnetaan. Käyttämällä kielimallin eri sanajonoille antamia prioritodennäköisyyksiä, voidaan tunnistuksen aikana ratkaisevasti rajoittaa läpikäytävien mahdollisten sanayhdistelmien määrää sekä erotella toisistaan homonyymit. Yksinkertainen laajan sanaston tunnistuksessa käytetty matemaattinen malli puhutulle kielelle on n–gram, jossa jokaisella sanayhdistelmällä on tietty todennäköisyys. N–gram mallin avulla jokaisen sanasekvenssissä $W = w_1, w_2, \dots, w_K$ esiintyvän sanan w_k todennäköisyys on laskettavissa, riippuen $n-1$ edellisestä sanasta $w_{k-1}, \dots, w_{k-n+1}$:

$$P(w_k|w_{k-1}, w_{k-2}, \dots, w_1) = P(w_k|w_{k-1}, \dots, w_{k-n+1}) \tag{8}$$

Koko sanasekvenssin $W = w_1, w_2, \dots, w_K$ todennäköisyydeksi muodostuu puheentunnistuksessa yleisesti käytetyn 3–grammin (trigram) avulla:

$$P(W) = P(w_1)P(w_2|w_1)\prod_{k=3}^{K} P(w_k|w_{k-1}, w_{k-2}), \tag{9}$$

jossa seuraavan sanan todennäköisyys riippuu kahdesta edellisestä sanasta. Trigrammi on erityisen tehokas, koska useimmilla sanoilla on vahva riippuvuus kahdesta edellisestä sanasta. Käytännössä kaikille harvinaisille sanayhdistelmille eli n–grammeille ei ole mahdollista, tai edes hyödyllistä estimoida omia trigrammi todennäköisyyksiä, vaan näiden sanojen kohdalla sovelletaan $n-1$ (unigram) tai $n-2$ (bigram) todennäköisyyksiä. (Deng, L. & Huang, X. 2009; Kurimo, M. 2008)

3.4 Puheen dekoodaus

Puheentunnistimen dekoodausprosessin tehtävä on matemaattisesti määritettynä löytää todennäköisin sanajono, jonka vastaavat akustiset mallit parhaiten täsmäävät puhesignaalista irrotettua piirrevektori sekvenssiä $Y = y_1, y_2, \dots, y_T$, ja joiden sanayhdistelmien todennäköisyydet täsmäävät kielimallin antamien sanojen priori todennäköisyyksiä. Tästä on edelleen johdettavissa todennäköisimpien HMM tilamallien hakutehtävä, joka hakee parhaan mahdollisen tilajonon leksikon ääntämismallien läpi. Tilajonoa vastaavan sanan prioritodennäköisyys voidaan ottaa haussa huomioon yksinkertaisemman kielimallin, kuten bigrammin avulla. Koulutettuja akustisia- ja kielimalleja sisältävää dekoodausprosessia kutsutaankin täten usein hakuprosessiksi. Käytännössä laajan sanaston jatkuvan puheentunnistuksessa todennäköisimpien tilajonojen haussa ei harkita kaikkia leksikon sanoja mahdollisena viestihypoteesina. Sen sijaan kaikki matala todennäköisyyksiset polut karsitaan pois mahdollisimman aikaisessa vaiheessa, välttäen näin tehokkaasti turhaa laskentaa. Näiden epätodennäköisimpien polkujen karsinta toteutetaan yleensä Viterbi-beam hakualgoritmilla, joka laskee jokaisella ajanhetkellä kaikkein todennäköisimmän polun/tilan, jonka jälkeen kaikki tietyn kynnysarvon alle jäävät tilat karsitaan pois. Tästä saadun parhaan tunnistushypoteesin lisäksi tuotetaan usein myös lista seuraavaksi parhaista lausahdus hypoteeseista (N–best list) tai sanakaavio (word lattice), jossa kullakin hypoteesilla on oma akustinen todennäköisyys ja kielimallin bigrammin priori todennäköisyys. Näiden parhaimpien lausahdus hypoteesien todennäköisyydet voidaan nyt pisteyttää uudelleen monimutkaisempien kielimallien, kuten trigrammin avulla. Tämän monimutkaisemman kielimallin tuottamilla priori todennäköisyyksillä korvataan jokaisen hypoteettisen lausahduksen aiemmat bigrammi todennäköisyydet uusilla trigram todennäköisyyksillä. Tästä saatujen uudelleen pisteytettyjen hypoteettisten lausahduksien lopputulemana valitaan se lausahdus, jonka akustinen ja

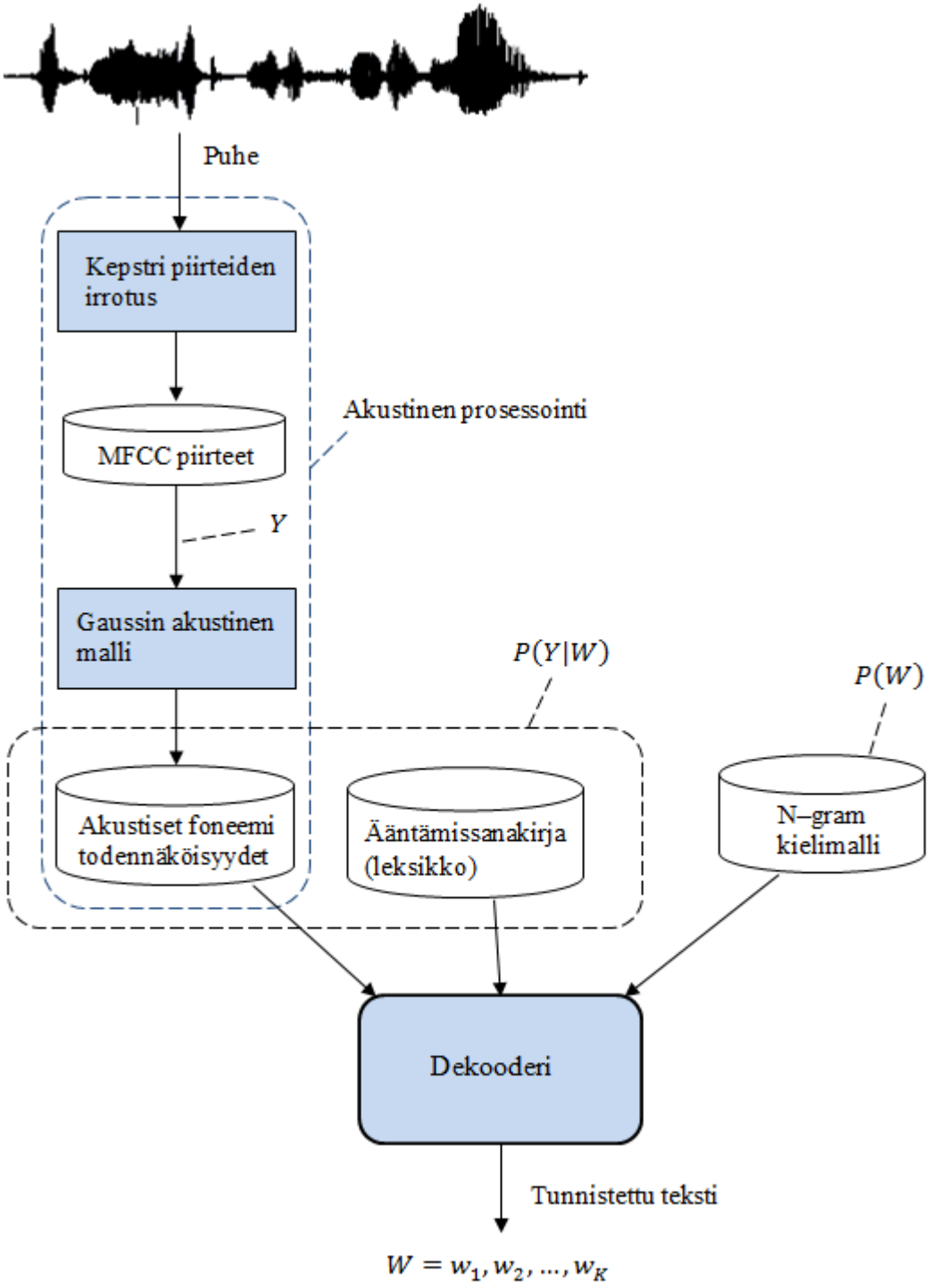
kielimallin todennäköisyys on kaikkein suurin. (Jurafsky, D. & Martin, J.H. 2008; Kurimo, M. 2009)

Yhteenvetona esitettynä, todennäköisin puheen sisältöä vastaava sanajono $W = w_1, w_2, \dots, w_K$, saadaan yhdistämällä dekooderin avulla jokaisen lausahduksen akustinen todennäköisyys sekä sanojen ja sanajonojen priori todennäköisyydet, ja valitsemalla potentiaalisista vaihtoehdoista kaikkein todennäköisin lausahdus. Matemaattisesti tehtävää kuvataan usein Bayesin säännön avulla, jonka lopputuloksena paras mahdollinen sanajono on se, joka maksimoi kielimallin priori ja akustisen mallin akustisen todennäköisyyden:

$$\widehat{W} = arg \max_{w}[P(Y|W)P(W)], \tag{10}$$

jossa akustinen todennäköisyys $P(Y|W)$ lasketaan akustisessa mallissa sovittamalla puhesignaalia äännemalleihin ja vaihtoehtoisiin sanahypoteeseihin, kun taas sanahypoteesin priori todennäköisyys $P(W)$ saadaan kielimallista edellä mainitun n–grammin avulla. Akustisessa mallissa laskettu ehdollinen todennäköisyys $P(Y|W)$, ilmaisee todennäköisyyden jolla sanajono W tuottaa piirrevektoreista Y havaittujen foneemien tilajonon ja kielimallissa laskettu todennäköisyys $P(W)$ ilmaisee sanojen priori todennäköisyyden riippuen edellisistä sanoista, siis riippumatta mitatusta signaalista. (Jurafsky, D. & Martin, J.H. 2008; Kurimo, M. 2008)

Mahdolliset foneemiyhdistelmät on listattuna tunnistimen leksikossa, jossa on kaikki sanat, jotka tunnistin tuntee ja niihin liittyvät ääntämistavat, jokainen ääntämistapa on esitettynä foneemijonona. Jokaista sanaa voidaan täten ajatella HMM ketjuna, jossa foneemit (tai sen rakenneosat) ovat HMM tiloja, ja Gaussin todennäköisyys estimaattorit antaa jokaisen HMM tilan ulostulon todennäköisyyden. Kuviossa 13 on esitettynä HMM–pohjaisen jatkuvan puheentunnistusjärjestelmän rakenne ja toiminta, jossa dekooderi yhdistää akustisessa mallissa Mel–kepstri-piirteistä havaitut foneemi todennäköisyydet, ja kielimallin antamien sanojen/sanajonojen todennäköisyydet, josta saadaan ulostulona kaikkein todennäköisin sanajono.



KUVIO 13. HMM–pohjaisen laajan sanaston jatkuvan puheentunnistusjärjestelmän (LVCSR) toimintakaavio. (Jurafsky, D. & Martin, J.H. 2008)

4 YHTEENVETO JA TULEVAISUUDEN SUUNTA

Puheentunnistuksella on ollut pitkä kehityksen historia, mutta vasta tilastollisten lähestymistavan myötä tutkimusala on ollut vakaassa kehityksessä ja avannut useita käytännön sovellusalueita. Erityisesti mobiililaitteiden yleistymisen myötä puheella ohjattavien käyttöliittymien kehitys on noussut merkittävästi ja siten nostanut puheteknologian merkittäväksi tutkimusalaksi. Tämä on lisännyt puheentunnistustutkimuksen käytössä olevia resursseja ja potentiaalista taloudellista merkitystä viime vuosikymmenen aikana huomattavasti. Etenkin tulevaisuudessa älykkään robotiikan astuessa mukaan työelämään, käyttäjäystävällinen puheentunnistus tulee kasvamaan yhteiskunnallisesti merkittäväksi teknologiaksi.

Useimmat nykyisistä puheentunnistusjärjestelmistä käyttävät Markovin piilomalleja (HMM) käsittelemään puheen ajallista vaihtelevuutta ja Gaussin mikstuurimalleja (GMM) määrittämään kuinka hyvin jokaisen Markovin piilomallin tila sopii puhekehyksen kertoimiin. Viime vuosina nopea kehitys koneoppimisen algoritmeissa ja koneiden laskentatehossa on johtanut vaihtoehtoiseen tapaan määrittämään tilojen sopivuutta käyttämällä monipiilokerroksisia syviä neuroverkkoja (DNN), jotka ottavat sisääntulona useita piirrekehyksiä, ja tuottavat piilokerrosten läpi posteriori todennäköisyydet HMM tiloille ulostulona. Syvien neuroverkkojen ja Markovin piilomallien DNN–HMM hybridi malli on viime vuosina osoittanut, että syvät neuroverkot voivat suoriutua Gaussin mikstuurimalleja paremmin akustisessa mallinnuksessa useissa eri puheentunnistuksen suorituskykytestissä.

Tällä hetkellä suurin haittapuoli syvissä neuroverkoissa verrattuna Gaussin mikstuureihin on, että niiden koulutus massiivisesta aineistosta on paljon haastavampaa. Tätä ongelmaa kompensoi hieman se, että neuroverkot käyttävät aineistoa paljon tehokkaammin hyödyksi, jolloin ne eivät vaadi yhtä paljon koulutus aineistoa saavuttaakseen saman suorituskyvyn. Ratkaisemalla koulutukseen liittyvät ongelmat, syvät neuroverkot uusilla oppimisalgoritmeilla tulevat lähitulevaisuudessa antamaan huomattavasti paremman puhujasta riippumattoman puheentunnistuksen.

LÄHTEET

Benesty, J.; Huang, Y. & Sondhi, M. (2008). *Springer handbook of speech processing* (ss. 539–549). Berlin: Springer.

Deng, L. & Huang, X. (2009). *An Overview of Modern Speech Recognition*. Haettu 21.12.2014 osoitteesta http://research.microsoft.com/pubs/118769/Book-Chap-HuangDeng2010.pdf

Gales, M. & Young, S. (2008) *The Application of Hidden Markov Models in Speech Recognition*. Haettu 23.12.2014 osoitteesta http://www.cslu.ogi.edu/~zak/cs506-lvr/mjfg_NOW.pdf

Giampiero, S. *Developing acoustics models for automatic speech recognition*. Haettu 28.12.2014 osoitteesta http://www.speech.kth.se/prod/publications/files/1308.pdf

Gmoore. (2005) *How Speech Recognition Works*. Haettu 24.5.2015 osoitteesta http://www.extremetech.com/computing/75394-how-speech-recognition-works8212and-doesnt-work/1

Jurafsky, D. & Martin, J.H. (2008). *SPEECH and LANGUAGE PROCESSING; An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 2. Painos. Haettu 26.12.2014 osoitteesta http://stp.lingfil.uu.se/~santinim/ml/2014/JurafskyMartinSpeechAndLanguageProcessing2ed_draft%202007.pdf

Kurimo, M. (2008). *Puhe ja kieli 28:2; Puheentunnistus* (ss. 73–83). Haettu 2.4.2015 osoitteesta http://ojs.tsv.fi/index.php/pk/article/view/5112/4616

Kurimo, M. (2009). Teoksessa *Puhuva ihminen: puhetieteiden perusteet* (ss. 336–342). Helsinki: Otava.

Kevin, M. (2008). *Estimation of Cepstral Coefficients for Robust Speech Recognition*. Haettu 2.4.2015 http://povinelli.eece.mu.edu/publications/papers/indrebophd.pdf

Rosti, A.-V. (2004). *Linear Gaussian Models for Speech Recognition*. Haettu 2.4.2015 osoitteesta http://mi.eng.cam.ac.uk/~mjfg/thesis_avir2.pdf

Sadewo, B. (2012). *Speech recognition: life before Siri, and what's to come.* Haettu 12.8.2015 http://www.androidauthority.com/speech-recognition-life-before-siri-and-whats-to-come-67994/

Ursin, M. (2002). *Triphone clustering in Finnish continuous speech recognition*. Haettu 1.6.2015 osoitteesta http://research.spa.aalto.fi/publications/theses/ursin_mst.pdf