Äännemallien vertailua jatkuvassa suuren sanaston puheentunnistuksessa

Vesa Siivola, Teemu Hirsimäki ja Mikko Kurimo
Teknillinen korkeakoulu
Informaatiotekniikan laboratorio
Vesa.Siivola@hut.fi, Teemu.Hirsimaki@hut.fi, Mikko.Kurimo@hut.fi

Tiivistelmä

Tässä työssä on verrattu erilaisia äännemalleja jatkuvassa suomenkielisessä puheentunnistuksessa. Käytetyt mallit ovat tavallinen foneemimalli (monofonimalli), kontekstiherkkä foneemimalli (trifonimalli) ja tavumalli. Lähtökohtana on foneemimalli, jossa jokaiselle suomen kielen foneemille on oma akustinen mallinsa. Käytännössä foneemin ääntämiseen vaikuttavat merkittävästi sitä ympäröivät foneemit, joten tarkempaa mallinnusta varten tehdään joka foneemille useita kontekstiriippuvia malleja. Tällainen trifonimalli on käytössä useimmissa moderneissa puheentunnistimissa. Suomen kielessä on lisäksi selkeä tavurakenne, joka on kuultavissa foneemien painotuksessa. Tässä työssä rakennettiin joka tavulle oma mallinsa ja kokeiltiin, miten tällainen malli toimii. Malleja verrattiin sekä jatkuvasta puheesta poimittujen yksittäisten sanojen että jatkuvan puheen tunnistuksessa. Jatkuvan puheen testit tehtiin sekä kielimallin kanssa että ilman. Puheaineistona käytettiin ääneen luettua kirjaa. Havaittiin että trifonimalli on vertailluista malleista paras.

1. JOHDANTO

Automaattinen puheentunnistus perustuu siihen, että puheesta voidaan irrottaa tunnistettavia tapahtumia. Tapahtuman tulee olla aina esiintyessään samankaltainen aikaisemmin havaittujen saman luokan tapahtumien kanssa, mutta riittävän erilainen kaikista muista tapahtumista. Eräs tällainen tapahtuma on sana. Sanoihin perustuvan tunnistimen opetuksessa tulee kuitenkin suuria käytännön ongelmia: opetusjoukossa ei voi etenkään suomen kielessä esiintyä kaikkia tarpeellisia sanoja. Lisäksi sopivan tilastollisen mallin opettaminen vaatisi useita opetusnäytteitä kutakin sanaa kohden. Malleista tulisi hyvin suuria, ja tunnistamiseen tarvittava laskenta-aika kasvaisi liian suureksi. Tämän takia sanatunnistimia käytetään vain pienehköissä tunnistimissa, jotka pystyvät tunnistamaan muutamia satoja avainsanoja [1].

Sana voidaan hajottaa pienempiin yksiköihin. Luonnollinen valinta on hajoittaa sana foneemeihin. Kirjoitettu suomen kieli on helppo muuttaa automaattisesti foneettiseen muotoon, pitää vain ottaa huomioon äng-äänne ja foneemien keston vaihtelu (TULI-TUULI). Hajottamalla sana foneemeiksi vältetään sanamallin pahimmat ongelmat. Joudutaan opettaa vain muutamia kymmeniä malleja (esim. 26 kpl), joten jo muutamassa tunnissa puhetta on riittävästi aineistoa hankalammankin tilastomallin opetukseen ainakin yhdelle puhujalle vakio-olosuhteissa. Nyt törmätään kuitenkin toisenlaiseen ongelmaan: foneemimallit eivät erota oikeita fooneja vääristä kovin hyvin, koska samaa foneеmia vastaava äänne (fooni) voi kuulostaa hyvin erilaiselta eri tilanteissa. Tähän vaikuttaa koartikulaatio, painotus ja intonaatio. Puheessa foneemi muuttuu liukuvasti toiseksi (ihmisen ääntöelimet eivät ole rajattoman nopeita), joten ympäröivät foneemit vaikuttavat lausuttavan foneemin ääntöön. Suomen kielessä paino on yleensä ensimmäisellä tavulla sekä tavujen alussa, joten nämä foneemit lausutaan eri tavalla kuin painottomat foneemit.

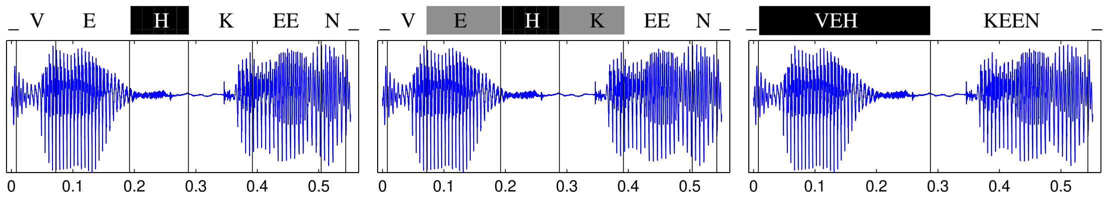
Voimme ottaa koartikulaation huomioon vaihtamalla tunnistuksen perusyksiköksi kontekstistaan riippuvan foneemin [2]. Tavupainotus voidaan huomioida käyttämällä joka tavulle omaa akustista malliaan [3]. Näiden mallien kanssa törmätään kuitenkin helposti samaan ongelmaan kuin kokonaisten sanojen kanssa: kaikille yksiköille ei enää löydy opetusaineistoa. Esimerkiksi trifonimallin opetukseen tarvittaisiin kaikille $26^3 \approx 17500$:lle eri mallille riittävä määrä näytteitä tilastollisen mallin opettamiseen. Joudutaan siis tasapainoilemaan mallin kompleksisuuden ja tämän kompleksisuuden hallinnan kanssa. Annetulla datamäärällä voidaan opettaa vain tietty määrä parametreja, jotka tulisi käyttää mahdollisimman tehokkaasti.

2. MALLIT

2.1. Rakenne

Foneemimalli

Foneemimallissa oletetaan, että puheesta voidaan löytää foneemirajat, joiden kohdalta sanat voidaan leikata paloiksi. Saman foneemin eri toteutuneet äänteet (foonit) oletetaan keskenään samankaltaisiksi, mutta erilaisiksi eri foneemeiden foonien kanssa (Kuva 1a).



(a) Monofonimalli (b) Trifonimalli (c) Tavumalli

Kuva 1: *Malli käyttää piirteet mustalla merkityltä alueelta. Käytetty malli valitaan harmaalla merkityn kontekstin perusteella.*

Trifonimalli

Trifonimallin toinen nimi, kontekstiriippuva foneemimalli [2], kuvaa menetelmää paremmin. Siinäkin puhe leikataan foneemin kokoisiin paloihin. Kullekin foneemille rakennetaan useita malleja kontekstin määrätessä, mikä näistä malleista valitaan (Kuva 1b). Trifonimalli parantaa monofonimallia ottamalla huomioon viereisten foneemien aiheuttaman koartikulaation.

Tavumalli

Tavumallissa [3] puhe jaetaan tavun kokoisiin osiin (Kuva 1c). Tavumalli parantaa monofonimallia ottamalla huomioon tavun sisäiset koartikulaatiot ja tavun painotuksen. Malli jättää kuitenkin huomiotta tavujen välisen koartikulaation.

2.2. Käytännön ongelmat

Ongelma, joka ei paljastu ennenkuin tarkastellaan mallien perustana olevaa kätkettyä Markov-mallia tarkemmin, on suomen kielen äännekeston mallinnus. Kätkettyä Markov-mallia käytetään, koska sillä pystytään luomaan monipuolisia malleja ja käsittelemään niitä tehokkaasti. Nämä mallit ovat kuitenkin huonoja erottamaan pitkiä ja lyhyitä foneemeja. Laboratoriomme vanhassa puheentunnistimessa [4] kokeiltiin menetelmää äännepituuksien erottamiseen oppivan vektorikvantisaation avulla [5], mutta uusimpaan tunnistimeen tätä ominaisuutta ei ole vielä siirretty. Näin ollen siis akustiset mallit eivät pysty erottamaan sanoja TUULA-TULLA-TUULLA toisistaan. Tähän asiaan palataan vielä pohdinnan yhteydessä.

Foneemimalli

Foneemimallissa ei ole opetukseen liittyviä ongelmia; se vain ei ole riittävän hyvä puheen tunnistamiseen [6].

Trifonimalli

Jos jokaiselle foneemille luodaan erillinen malli kaikille konteksteille, saadaan kymmeniä tuhansia äännemalleja. Tätä määrää varten ei yleensä pystytä löytämään riittävästi opetusaineistoa, joten mallien määrää pitää vähentää. Tämän voi tehdä ainakin kahdella tavalla:

1) Opetetaan vain ne mallit, joille opetusaineisto riittää Tässä valittiin ne trigrammit, jotka esiintyivät yli 50 kertaa. Jäljellä olevista niille, joille data riittää jomman kumman kontekstin opetukseen, opetetaan tällainen yksipuolinen kontekstimalli (difonimalli). Niille, joille data ei riitä tähänkään, opetetaan kontekstiriippumaton foneemimalli.

2) Yhdistetään samankaltaiset mallit yhdeksi malliksi. Esimerkiksi A:n mallit P+A+U ja T+A+U voisivat olla riittävän samanlaisia, että ne voitaisiin yhdistää.

Olemme tässä käyttäneet yksinkertaisuuden vuoksi ensimmäistä vaihtoehtoa, vaikkakin toisella on saatu erinomaisia tuloksia [7].

Tavumalli

Tavumallissa on sama ongelma kuin trifonimallissa: mallien määrä on liian suuri opetettavaksi. Tämä voidaan kiertää opettamalla mallit yleisimmille tavuille ja rakentamalla muiden tavujen mallit vastaavista monofonimalleista.

Yksityiskohtana voidaan mainita, että aivan koulussa opetetulla tavutuksella ei kannata malleja rakentaa. Esimerkiksi hajottamalla sana TAKKA osiin TAK ja KA ei saada akustisesti järkevää mallia, sillä K:n “kolahdus” tapahtuu vain kerran, jälkimmäisessä tavussa. Näin esimerkiksi sanasta TAK-SI opetettu akustinen malli tavulle TAK ei vastaisi lainkaan sanasta TAK-KA opetettua mallia. Sanan TAKKA oikeampi jako olisi TA-KKA. Nyt tavu TA vastaa esimerkiksi sanasta TA-RU opetettua mallia.

2.3. Kielimalli

Käytetty kielimalli oli englanninkielisissä tunnistimissa yleisesti käytetty trigrammimalli. Siinä uuden sanan todennäköisyys riippuu kahdesta edellisestä sanasta eli malli antaa todennäköisyyden $P(w_3|w_2, w_1)$. Kielimallin sopivuutta tekstiaineistoon voidaan mitata laskemalla mallin antaman todennäköisyyden geometrinen keskiarvo (hämmentyneisyys, perplexity):

$$Perp(w_1, \ldots, w_N|M) = P(w_1, \ldots, w_N|M)^{-\frac{1}{N}}, \quad (1)$$

missä M on annettu malli.

3. KOKEET

Koejärjestelyissä pyrittiin siihen, että tulokset vastaisivat mahdollisimman tarkasti tilannetta, jossa lukijan äänelle opetettua tunnistinta käytetään uuden aineiston tunnistamiseen.

3.1. Aineisto

Akustinen aineisto

Akustisena aineistona käytettiin naislukijan ääneen lukemaa kirjaa "Syntymättömien sukupolvien Eurooppa" [8]. Siitä ensimmäiset 12 tuntia (49 000 sanaa) käytettiin akustisten mallien opetukseen ja loput, noin 45 minuuttia (3 000 sanaa), käytettiin testaamiseen. Aineistosta opetettiin 26 monofonimallia, 413 tavumallia tai 1289 trifonimallia. Monofonimallissa oli noin 71 000 parametria, tavumallissa 286 000 ja trifonimallissa 431 000.

Tekstiaineisto

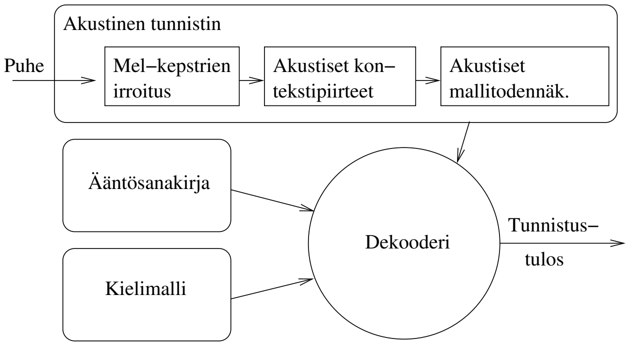
Kielimallin ja sanaston opettamiseen käytettiin akustisesta aineistosta riippumatonta aineistoa, CSC:n Kielipankkia [9] ja STT:n sähkeuutisia. Näiden pituus yhteensä oli noin 40 miljoonaa sanaa. Aineistosta etsittiin 64 000 yleisintä sanaa, joille opetettiin trigrammimalli. Silti 20% testiaineston sanoista ei löytynyt sanastosta.

3.2. Ohjelmistot

Suurin osa kokeisiin liittyvästä ohjelmistosta on toteutettu itse. Suomen kielen tavutus perustuu yksinkertaisiin sääntöihin, poikkeuksena yhdyssanat, joiden kohdalla pitäisi pystyä erottelemaan sanan merkitykset. Tekemämme tavutin siis tavuttaa oikein kaiken paitsi osan yhdyssanoista, joihin sen yksinkertainen säännöstä ei riitä. Vääriä tavutuksia oli kuitenkin vähän.

Kokeessa käytetty puheentunnistin (Kuva 2) on kaikilta osiltaan tehty itse. Akustinen osa perustuu kätkettyihin Markov-malleihin [10], joiden emissiotodennäköisyydet estimoitiin Gaussisena mikstuurimallina. Mikstuurimalli alustettiin [4] itseorganisoituvalla kartalla [11] ja opetettiin Viterbi-opetuksella [12]. Piirteinä käytettiin mel-kepstrejä [13] 16 ms ikkunalla. Monofoneissa ja kontekstiriippuvissa foneemeissa on 3 tilaa, joissa kussakin on itsetransitio ja transitio seuraavaan tilaan. Tavumalleissa tilojen määrä on $2n + 1$, missä n on tavun foneemien määrä.

Käytetty dekooderi [14] on ns. pinodekooderi, jossa samoihin ajanhetkiin loppuvat tunnistushypoteesit kerätään omiin pinoihinsa. Samassa pinossa olevia hypoteeseja lavenetaan sana kerrallaan etsimällä kätketyn Markov-malliverkon avulla akustisesti parhaiten



Kuva 2: *Puheentunnistusjärjestelmä*

sopivat sanat halutusta ajanhetkestä alkaen ja sijoittamalla lavennetut hypoteesit parhaan loppumisajan mukaisiin pinoihin. Kielimallia ei ole rakennettu kiinteästi verkkoon, vaan kielimallin antamat todennäköisyydet otetaan huomioon vasta siinä vaiheessa, kun hypoteesit sijoitetaan pinoihin. Tämä mahdollistaa mielivaltaisen pitkien kielimallien käytön.

Trigrammikielimallin opettamiseen käytettiin CMU-Cambridge Statistical Language Modeling Toolkit -ohjelmistoa [15].

3.3. Koejärjestelyt ja tulokset

Koska 20% testiaineiston sanoista puuttuu sanastosta, on tunnistimen mahdotonta tunnistaa näitä sanoja oikein. Lisäksi käytetyt mallit eivät pysty erottamaan foneemeja pituuden perusteella. Siksi tulokset annettiin sekä todelliselle tapaukselle että sellaiselle tapaukselle, jossa sanat oli muunnettu sisältämään vain lyhyitä foneemeja (eli sanat TUULA, TULLA ja TUULLA vastaavat kaikki tapahtumaa TULA). Näiden tulosten lisäksi kussakin on annettu optimistisen tunnistuksen tulos, missä sanastossa on vain 17 911 sanaa ja sanasto sisältää kaikki akustisen aineiston sanat. Nämä tulokset on laskettu vain trifonimallille huomioimatta foneemien pituuksia.

Jatkuvalla puheella käytettiin virhemittana sanavirheiden osuutta, oikein olevien sanojen osuutta ja foneemivirhettä.

$$\text{Sanavirhe\%} = \frac{\textit{lisäykset+korvaukset+poistot}}{\textit{alkuperäisten sanojen määrä}} \cdot 100\% \qquad (2)$$

$$\text{Oikein\%} = \frac{\textit{korvaukset+poistot}}{\textit{alkuperäisten sanojen määrä}} \cdot 100\% \qquad (3)$$

Foneemivirhe lasketaan kuin sanavirhe, mutta sanojen sijasta käytetään foneemeja.

Leikatut sanat

Ensimmäinen koe on helpotettu jatkuvan puheen tunnistamisen koe. Tässä tunnistimelle on kerrottu, mistä sana alkaa ja minne se päättyy. Tulokset on esitetty taulukossa 1. Optimistisella sanastolla trifonimalli teki 8% virheitä.

Jatkuva puhe ilman kielimallia

Toinen koe oli sama kuin ensimmäinenkin, paitsi että tunnistimelle ei kerrottu sanarajoja eikä sanojen määrää. Tulokset on esitetty taulukon 2 alussa. Optimistisella sanastolla trifonimallin virheprosentti oli 40% (foneemivirheprosentti 10%).

Malli	Foneemin pituus	
	ei huomioida	huomioidaan
Monofonit	38.2%	47.2%
Trifonit	29.6%	40.5%
Tavumallit	33.9%	44.2%

Taulukko 1: *Virheellisten sanojen osuus leikattujen sanojen testissä. Paras mahdollinen tulos on 20%, sillä tunnistimen sanasto ei sisällä kaikkia testijoukon sanoja.*

malli	foneemin pituus huomioitu		fon. pituus ei huomioitu	
	oikein%	sanavirhe	sanavirhe	fon.virhe
Ei kielimallia				
Monofonit	13%	**232**%	230%	49%
Tavumallit	24%	**214**%	210%	50%
Trifonit	46%	**111**%	105%	24%
Kielimallin kanssa				
Monofonit	29%	**106**%	104%	33%
Tavumallit	40%	**104**%	102%	32%
Trifonit	52%	**80**%	76%	19%

Taulukko 2: *Jatkuva puhe. Testijoukon sanoista 20% ei ollut mahdollista tunnistaa, sillä ne eivät olleet sanastossa.*

Jatkuva puhe

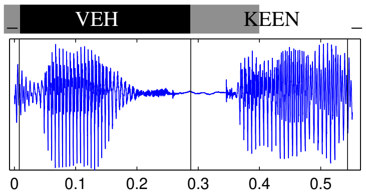
Viimeisessä kokeessa edellisen kokeen tunnistimen avuksi otettiin kielimalli. Tässä siis painotettiin kunkin sanan akustista todennäköisyyttä sanan kielimallitodennäköisyydellä ja saatiin näin parempi tulos. Testissä käytetyn trigrammikielimallin hämmentyneisyys (perplexity) testiaineistolla oli 4200. Tulokset on esitetty taulukon 2 loppupuolella. Optimistiselle sanastolle ei voitu opettaa järkevää kielimallia.

3.4. Tulosten tarkastelu

Trifonimalli on selvästi parempi kuin tavumalli. Sen sisältämä konteksti-informaatio on ilmeisesti tärkeämpää kuin tavujen sisäinen konteksti-informaatio ja niihin liittyvä painotus yhteensä. Huomattavan suuri osa jatkuvan puheen tunnistusvirheistä näytti johtuvan siitä, että oikean sanan puuttuessa sanastosta tunnistin sovittaa monta pikkusanaa tuntemattoman sanan kohdalle. Kielimalli karsii näitä lisäysvirheitä ja vähentää myös korvausvirheiden määrää. Poistovirheitä oli niin vähän, ettei niillä ole merkitystä.

4. POHDINTA

Jos trifonimallin ja tavumallin hyvät puolet haluaa yhdistää, saadaan kontekstiriippuva tavumalli (Kuva 3). Tässä mallissa kuitenkin trifonimallien ja tavumallien opetusvaikeudet kasautuvat, sillä malleja tulee liikaa opetettavaksi. Opetettavien parametrien määrää voidaan laskea pakottamalla eri malleille yhteisiä parametreja. Esimerkiksi voidaan rakentaa yhteinen viimeinen tila kaikille tavumalleille, jotka päättyvät H:hon ja joita seuraa K. Suurempi opetusjoukko tietysti auttaisi tällaisen mallin opetusta.



Kuva 3: *Kontekstiriippuva tavumalli*

Eräs mallin heikkous on, etteivät käytetyt mallit kykene erottamaan lyhyttä ja pitkää foneemia. Etenkin helpossa testissä (leikatut sanat) tämä heikkous näkyy selvästi. Kuitenkin realistisessa tapauksessa (jatkuva puhe kielimallin kanssa) foneemin pituuden tunnistamisen osuus virheestä on vielä hyvin pieni. Vaikka tämäkin ongelma on ratkaistava, toistaiseksi löytyy vielä paljon suurempiakin virhelähteitä, joiden ongelmat pitäisi ensin ratkoa.

Kielimallin malliyksikön koko

Perinteisissä englannin kielen tunnistimissa kielimallin mallinnusyksikkö on sana. Tämä toimii hyvin englannin kielellä, mutta suomen kielessä tämä valinta johtaa suuriin vaikeuksiin. Vaikka tunnistimen sanasto (64 000 sanaa) on suuri, se ei riitä lainkaan kattamaan testijoukon sanastoa: 20% sanoista on tuntemattomia. Tämä johtuu suomen kielen sanojen taipumisesta: yhdellä sanalla voi olla yleisessäkin käytössä satoja erilaisia tavutusmuotoja. Tämä sanojen taivutusmekanismi pitäisi pystyä kuvaamaan tilastollisin käsittein. Nyt sanojen välistä suhdetta kuvaava malli ei kuitenkaan enää toimi sellaisenaan, sillä trigrammimallin yksikkö on sana. Sanojen välinen suhde pitäisi mallintaa jollain muulla tavalla.

5. KIITOKSET

Kiitämme Näkövammaisten Keskusliittoa äänikirjamateriaalista ja Mietta Lennestä, Nicholas Volkia ja muita Helsingin yliopiston fonetiikan ja yleisen kielitieteen laitoksen tutkijoita, jotka auttoivat puheäänitteiden järjestelyssä ja äänikirjojen tekstien esikäsittelyssä. Uutisähkeet on hankittu Suomen Tietotoimistolta (STT), ja muuta tekstiaineistoa on käytetty CSC:n (Tieteellinen laskenta Oy) kielipankin kautta. Tutkimusta on osittain rahoittanut TEKESin USIX tutkimusohjelma. Kiitämme avusta myös projektimme Speech-To-Text partnereita.

6. VIITTEET

[1] L. Bahl, P. Brown, P. de Souza ja R Mercer, "Acoustic Markov models used in the Tangora speech recognition system", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1988, s. 497–500.

[2] R.M. Schwartz, Y. Chow, O. Kimball, S. Roucos, M. Krasner ja J. Makhoul, "Context-dependent modeling of acoustic-phonetic recognition of continuous speech", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1985, s. 31.3.1 – 31.3.4.

[3] J. Baker, “Syllable as a unit of speech recognition”, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no. 1, s. 82–87, 1975.

[4] M. Kurimo, *Using Self-Organizing Maps and Learning Vector Quantization for Mixture Density Hidden Markov Models*, Väitöskirja, Teknillinen korkeakoulu, 1997.

[5] Janne Jalkanen, “Foneemien pituuksien erottaminen”, erikoistyö Informaatiotekniikan laboratoriolle, Teknillinen korkeakoulu, 1997.

[6] Vesa Siivola, Mikko Kurimo ja Krista Lagus, “Large vocabulary statistical language modeling for continuous speech recognition”, *Proceedings of 7th European Conference on Speech Communication and Technology*, 2001.

[7] M. Ursin, “Triphone clustering in Finnish continuous speech recognition”, Diplomityö, Teknillinen korkeakoulu, 2002.

[8] STT-Speech-To-Text: Phonemic Speech Recognizer for Finnish-USIX research project of Technology Development Centre of Finland TEKES, “Speech database (in Finnish)”, http://www.acoustics.hut.fi/projects/USIX_STT/arkisto.html, 2001.

[9] CSC-Tieteellinen laskenta Oy, “Finnish Language Text Bank: Corpora Books, Newspapers, Magazines and Other”, http://www.csc.fi/ kielipankki/, 2001.

[10] L. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition”, *Proceedings of the IEEE*, vol. 77, no. 2, s. 257–286, 1989.

[11] T. Kohonen, *Self-Organizing Maps*, Springer, Berlin, 2001, 3. painos

[12] G. Forney, “The Viterbi algorithm”, *Proceedings of the IEEE*, vol. 61, no. 3, s. 268–278, 1973.

[13] Y. Tohkura, “A weighted cepstral distant measure for speech recognition”, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 10, s. 1414–1422, 1987.

[14] T. Hirsimäki, “A decoder for large-vocabulary continuous speech recognition”, Diplomityö, Teknillinen korkeakoulu, 2002.

[15] P. Clarkson ja R. Rosenfeld, “Statistical language modeling using CMU-Cambridge toolkit”, *Proceedings of 5th European Conference on Speech Communication and Technology*, Syyskuu 1997, s. 2707–2710.