ESIMERKKIPOHJAINEN MELUISAN PUHEEN AUTOMAATTINEN TUNNISTUS

**Antti Hurmalainen, Tuomas Virtanen,
Jort Gemmeke, Katariina Mahkonen**

Signaalinkäsittelyn laitos
Tampereen teknillinen yliopisto
PL553, 33101 Tampere
antti.hurmalainen@tut.fi, tuomas.virtanen@tut.fi,
jgemmeke@amadana.nl, katariina.mahkonen@tut.fi

1 JOHDANTO

Koneellinen puheentunnistus on pitkään tutkittu ongelma, jolle löytyy lukuisia käyttötarkoituksia mm. automaattisen transkription, puheohjattujen laitteiden ja tietojärjestelmien käytön aloilta. Nykyiset toteutukset pystyvät lähinnä häiriöttömissä oloissa hyvään tunnistustarkkuuteen, mutta laatu laskee nopeasti taustamelun myötä. Koska monet sovellukset ovat tarkoitetut käytettäviksi vaihtelevissa ääniympäristöissä, järjestelmien melusietoisuutta tulisi saada parannettua merkittävästi käyttövarmuuden takaamiseksi. Myös puhujariippumattomuus on toivottava piirre yleiseen käyttöön tarkoitussa järjestelmässä.

Enemmistö nykyisistä puheentunnistimista perustuu lyhyiden aikakehysten ($\approx$ 25 ms) analysointiin käyttäen multinormaalijakaumiin pohjautuvia spektrimalleja. Pitkään jatkuneen kehityksen ansiosta mallit on saatu toimimaan tehokkaasti puheen sisällön tunnistamista ajatellen. Lähestymistavan ongelmana on, että taustamelu turmelee helposti yksittäisten kehysten puhespektrin jolloin tunnistus usein epäonnistuu. Ongelmaa on yritetty ratkoa erilaisilla suodatus- ja kompensaatiomenetelmillä, mutta näillä saavutettavat parannukset eivät vielä ole tyydyttävällä tasolla laajaa käyttöä ajatellen.

2 ESIMERKKIPOHJAINEN PUHEENTUNNISTUS

Puheentunnistuksen melusietoisuutta on mahdollista parantaa siirtymällä pidempään aikakontekstiin. Yksittäisen kehyksen luokittelun sijasta siirrytään useiden kehysten pituiseen aikaikkunaan. Puhe- ja melupiirteille olisi mahdollista luoda tällöinkin mallipohjaiset aika-taajuusjakaumat. Käytännössä kuitenkin hyviä tuloksia on saavutettu käyttämällä *ääniesimerkkejä* (engl. “exemplar”). Tällöin puhe- ja melumallin kantaelementteinä toimivat suoraan harjoitusmateriaalista poimitut spektrogrammilohkot. Osa näistä mallintaa puhdasta puhetta, osa pelkkää melua. Esiteltävässä järjestelmässä käytetään 23 Mel-taajuuskaistaa ja 10–30 aikakehystä. Koska siirtymä kehysten välillä on 10 millisekuntia, yksi elementti ilmentää aika-taajuustason äänitapahtumia noin 100–300 millisekunnin ajalta.

Esimerkkipohjaisen menetelmän keskeinen ajatus melusietoisuuden kannalta on se, että

havaittu spektrogrammi voidaan esittää aiemmin kerättyjen esimerkkien avulla. Koska ikkunaan sisältyy pitkiä, aikatason muutokset huomioivia piirteitä, voidaan tällaisia kokonaishahmoja löytää myös melun keskeltä. Tyypillinen äänneyhdistelmä on mahdollista havaita silloinkin, kun se hetkellisesti peittyy kokonaan häiriöiden alle. Lyhyen kontekstin tunnistimille tällainen tilanne tuottaisi suuria ongelmia.

Toinen olennainen ominaisuus kuvattavassa järjestelmässä on meluisan havainnon mallintaminen puheen ja melun summana. *Ei-negatiivisen matriisihajotelman* (non-negative matrix factorisation, NMF) avulla voidaan löytää esitys, jossa kokonaishavainto kootaan summaamalla kantaesimerkkien painokertoimia, *aktivaatioita*. Ei-negatiivisuus takaa, että esimerkkejä ainoastaan lisätään toisiinsa positiivisella painolla tai ei käytetä lainkaan. Kun piirteet ovat positiivisia spektrimagnitudeja, tämä vastaa läheisesti todellista, fyysistä mallia äänten summautumisesta.

Lisäehtona hajotelmassa käytetään *harvuutta* (sparsity). Algoritmia ohjataan siten, että havainto mallinnetaan harvojen kantaelementtien yhdistelmänä. Toisin sanoen aktiivisia lähteitä sallitaan vähemmän kuin täysin vapaassa hajotelmassa syntyisi. Tämä pakottaa järjestelmän löytämään sellaisia ratkaisuja, joissa pieni osa esimerkeistä riittää selittämään havainnon kohtalaisella tarkkuudella. Koska etenkin puheen kohdalla puhuja voi tuottaa vain yhtä äännettä kerrallaan, harva esitys tuottaa sille uskottavan mallin. Liian monien aktivaatioiden salliminen johtaisi ylisovittamiseen epärealistisen suurella määrällä eri lähteitä.

Kun havainnolle on löydetty esitys puhe- ja meluelementtien summana, voidaan esimerkiksi poistaa kokonaan meluaktivaatiot ja rekonstruoida signaali pelkästään havaittujen puheaktivaatioiden avulla. On tosin myös näytetty, että tämä vaihe ei ole tarpeen puheentunnistusta ajatellen. Puheen sisältö voidaan tulkita suoraan puheaktivaatioiden identiteetistä ja painoista. Tällä tavoin saadut tunnistustulokset ovat parempia kuin ehostuspohjaisilla, signaalia puhdistavilla menetelmillä on saavutettu. [1, 2, 3]

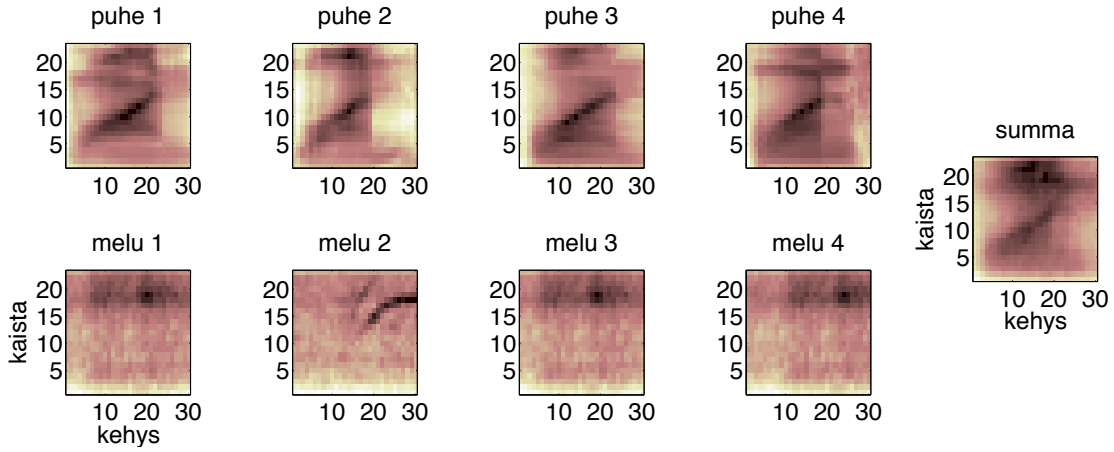
3 MATEMAATTINEN MALLI JA TUNNISTUSJÄRJESTELMÄ

Otetaan tutkittavaksi yksi havaintoikkuna, jossa on T peräkkäistä kehystä ja B taajuuskaistaa. Sen kaikki aika-taajuuselementit (spektrimagnitudit) voidaan esittää $T \cdot B$-pituisena vektorina $\mathbf{y}$. Jos sanakirja koostuu K samankokoisesta spektrogrammista, samoin piirrevektoreiksi muunnettuna $\mathbf{a}_i$ ($i \in 1 \ldots K$), havainnon painotettu summaesitys on

$$\mathbf{y} \approx \sum_{i=1}^{K} x_i \mathbf{a}_i. \tag{1}$$

Siispä painokertoimet x_i määrittävät sanakirjaelementtien suhteelliset osuudet havainnon selittämiseksi. Harvuus tarkoittaa, että enemmistö painoista on nollia tai hyvin pieniä. Kuvassa 1 on esitetty, kuinka havainto voi koostua useiden eri esimerkkien yhdistelmästä.

Kaikki summaan valikoituvat puhe-esimerkit ovat sanasta “one”. Vaikka mikään näistä



Kuva 1: *Piirresummaus*

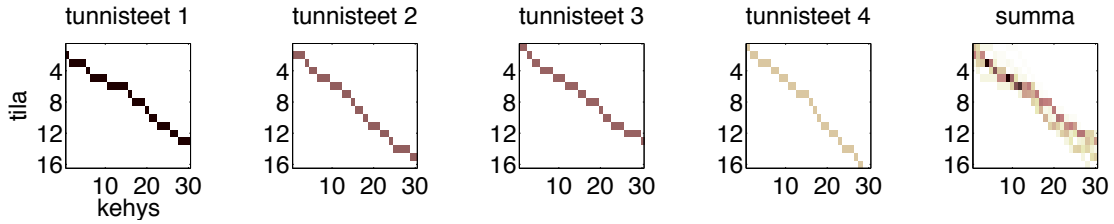
ei täsmälleen vastaa havaittua sanaa, yhdessä ne pystyvät esittämään hiukan erilaisenkin ääntämyksen. Lisäksi osa havainnosta selittyy melulla.

Ikkunan kaikki painot voidaan esittää yhdellä vektorilla $\mathbf{x}$. Kun sama menettely toistetaan peräkkäisille, lomitetuille aikaikkunoille, saadaan jokaiselle ikkunalle $j \in 1 \ldots L$ oma painovektori $\mathbf{x}_j$. Nämä yhdessä kertovat, mitkä kantaelementit aktivoituvat eri vaiheissa tunnistettavaa lausetta.

Lauseen sisältö voidaan selvittää suoraan aktivaatioista. Jokaisen sanakirjan esimerkin sisältö voidaan olettaa ennalta tunnetuksi. Niinpä puhetta vastaaviin esimerkkeihin voidaan sisällyttää *tunnisteet* (label), jotka kertovat niiden foneettisen sisällön.

Oletetaan, että järjestelmässä on Q foneettista tilaa. Esitetään jokaisen ajanhetken j tilasisältöä vektorilla $\mathbf{q}_j$, jonka pituus on Q. Jokainen vektorin elementti kertoo, kuinka todennäköinen kukin tila on kyseisellä ajanhetkellä. Kun tietty puhe-esimerkki aktivoituu hetkellä j, siihen kuuluvien tilojen todennäköisyyttä kasvatetaan aktivaatiota vastaavalla aikavälillä.

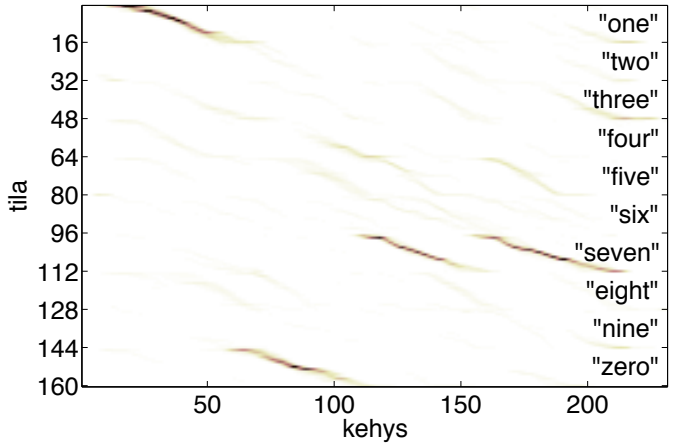
Kuvassa 2 on esitetty yllä nähtyjen puhe-esimerkkien tunnisteketjut ja niistä syntyvä yhdistelmä.



Kuva 2: *Tunnistesummaus*

Erilaiset sanan “one” esimerkit tuottavat hiukan toisistaan poikkeavat tilaketjut, mutta niiden yhdistelmänä saatava kuvio on helposti tunnistettavissa oikeaksi sanaksi. Sama toistetaan kaikkien aktivaatioiden kohdalla. Lopputuloksena saadaan tilamatriisi $\mathbf{Q}$, josta nähdään puheen sisältö pidemmältä ajalta. Kuvassa 3 nähdään kokonainen lauseen

tilamatriisi. Lauseen kesto on noin 260 kehystä (2.6 s) ja puhetilat esittävät erilaisia ääneen lausuttuja numeroita.



Kuva 3: *Tilamatriisi*

Kuvasta voidaan lukea, että lauseessa esiintyy englanninkielinen numerosarja “1077”. Tulkitseminen tapahtuu tehtävään sopivalla kielimallilla, kuten tavanomaisissakin järjestelmissä.

4 TULOKSIA

Esimerkkipohjaisen tunnistuksen tehokkuutta on testattu AURORA-2 -tietokannalla, jossa useat eri puhujat luettelevat englanninkielisiä numerosarjoja. Jokainen lause koostuu 1–7 numerosta, ja tehtävänä on havaita mahdollisimman moni numero oikein. Todellisten tunnistustilanteiden mallintamiseksi lauseisiin on lisätty taustamelua signaali-kohinatasoilla 20:stä -5 desibeliin. Toisin sanoen vaikeimmissa tapauksissa melu on kokonaisuutena voimakkaampaa kuin puhe itse. Melua on äänitetty kahdeksasta eri lähteestä (maanalainen, auto, puheensorina, näyttely, ravintola, katu, lentokenttä, rautatieasema).

Tunnistusjärjestelmää varten laadittiin sanakirjat, joihin valittiin 5000 esimerkkiä puheesta sekä toiset 5000 esimerkkiä melusta. Jälkimmäistä varten materiaalia oli saatavilla vain neljästä ensimmäisestä melulähteestä. Sanakirjat koottiin kolmella eri ikkunanpituudella: 10, 20 tai 30 peräkkäistä kehystä ikkunaa kohden. Sanakirjoihin ja havaintopiirteisiin sovellettiin taajuuskaistavoimakkuuksien tasausta. Tämän jälkeen lausepiirteistä laskettiin hajotelmat ja näin saadut aktivaatiot tunnistettiin edellä kuvatulla tavalla. Tulokset on koottu taulukoihin 1 (melujoukko ’A’) ja 2 (melujoukko ’B’). Luvut kuvaavat oikein tunnistettujen sanojen prosentteja eri kohinatasoilla ja ikkunanpituuksilla. Lisäksi on esitetty vertailutulokset tavanomaisemmasta, jonkin verran melua kompensoivasta järjestelmästä.

Joukon ’A’ taustamelu käsittää ne neljä lähdettä, joista oli saatavilla esimerkkejä sanakirjoja varten. Joukon ’B’ häiriö oli ennalta tuntematonta.

Taulukko 1: A-tuloksia

SNR	clean	20	15	10	5	0	-5
vert.	99.7	97.9	95.5	91.4	82.6	62.1	17.1
T=10	96.2	95.3	94.4	92.1	84.7	71.2	39.6
T=20	96.6	95.8	94.8	92.7	88.8	78.1	53.1
T=30	94.7	93.4	93.3	92.2	89.9	79.5	56.7

Taulukko 2: B-tuloksia

SNR	clean	20	15	10	5	0	-5
vert.	99.7	95.3	91.2	84.3	70.4	40.2	12.2
T=10	96.2	94.7	93.6	87.9	78.4	57.1	27.4
T=20	96.6	95.3	93.7	89.9	82.7	63.1	35.7
T=30	94.7	93.5	93.2	90.1	85.7	67.5	37.6

Nähdään, että normaalijakaumien mallinnukseen perustuva vertailujärjestelmä suoriutuu paremmin puhtaan puheen tunnistuksesta. Tässä suhteessa esitetty malli ei vielä ole optimoitujen tunnistimien tasolla. Toisaalta kohinan lisääntyessä esitetty algoritmi saavuttaa merkittävästi parempia tuloksia. -5 desibelin kohdalla vertailumenetelmä muuttuu käytännössä käyttökelvottomaksi, mutta esimerkkipohjainen tunnistin saavuttaa yhteensopivan sanakirjan kanssa yli 50 prosentin tunnistustarkkuuden.

5 DISKUSSIO

Esimerkkipohjainen puheentunnistusmenetelmä on osoittautunut perinteisiä normaalijakaumien mallinnukseen perustuvia menetelmiä toimivammaksi erittäin häiriöisessä ympäristössä. Ääniesimerkkien käyttö on kätevä tapa mallintaa äänneyhdistelmien laajempia hahmoja. Mitä enemmän signaalissa on häiriötä, sitä enemmän puhe-esimerkkien ajallisesta pituudesta on etua tunnistukseen. Häiriöosuus on helppo poistaa puheesta, joka on mallinnettu summautuvien puhe- ja meluesimerkkien avulla. Tällöin mallia voidaan käyttää myös puheäänen laadun parantamiseen.

VIITTEET

[1] VIRTANEN T, GEMMEKE J, & HURMALAINEN A, State-based labelling for a sparse representation of speech and its application to robust speech recognition, in *Proceedings of INTERSPEECH2010*, pages 893–896, Makuhari, Japan, 2010.

[2] RAJ B, VIRTANEN T, CHAUDHURE S, & SINGH R, Non-negative matrix factorization based compensation of music for automatic speech recognition, in *Proceedings of INTERSPEECH2010*, pages 717–720, Makuhari, Japan, 2010.

[3] GEMMEKE J, VIRTANEN T, & HURMALAINEN A, Exemplar-based sparse representations for noise robust automatic speech recognition, *IEEE Transactions on Audio, Speech and Language Processing*, (2011).