Pertti Väyrynen & Tapio Seppänen

Puheen prosodiset piirteet audiotallenteiden sisällön kuvailussa ja tiedonhaussa

Pertti Väyrynen & Tapio Seppänen: Puheen prosodiset piirteet audiotallenteiden sisällön kuvailussa ja tiedonhaussa [Utilising prosodic features of speech in content descriptions and retrieval of audio recordings] Informaatiotutkimus 21 (4), 83-89.

In this article, the possibility of utilising prosodic features of speech in Spoken Document Retrieval (SDR) is discussed. At present, it is traditional text-based retrieval techniques that are mainly exploited in SDR with good results. However, prosodic features of speech (non-lexical information) identifying, for example, the gender or age of the speaker offer new indexing features that are also potentially useful in SDR.

Address: Pertti Väyrynen & Tapio Seppänen, MediaTeam Oulu Group, University of Oulu, Erkki Koiso-Kanttilan katu 3, FIN-90014 University of Oulu, Finland
e-mail: pav@ee.oulu.fi.

1. Johdanto

Puheentunnistuksen virheistä johtuen audiotallenteiden kuten esim. spontaania puhetta sisältävien keskustelujen sisällönkuvailu esim. tavanomaisten indeksitermien avulla ja näiden indeksitermien ja dokumenttien sisällön väliseen vastaavuuteen perustuva tiedonhaku perinteisten tekstipohjaisten tiedonhakumenetelmien avulla ei todennäköisesti onnistu vielä lähitulevaisuudessa riittävän hyvin (Hindus et al. 1993) vaikka puheentunnistustekniikoiden suorituskyky paraneekin jatkuvasti. Ongelmana tämän tyyppisessä tiedonhaussa on virheellisesti tunnistettujen sanojen lisäksi myös puheentunnistusjärjestelmien liian suppea sanasto ja siitä johtuva kyvyttömyys tunnistaa erisnimiä (Ferrieux & Peillon 1999) tai spontaanin puheen ominaisuudet yleensä, mitkä vaikuttavat myös tunnistustarkkuuteen (Colineau & Halber 1999).

Tunnistusvirheistä kuten esim. väärin tunnistetuista sanoista huolimatta automaattista puheentunnistusta hyödynnetään audiotallenteiden sisällön kuvailussa ja tiedonhaussa perinteisten tekstipohjaisten tiedonhakumenetelmien avulla.

Puhekommunikaatiossa, joka sinällään on *multimodaalista* koostuen verbaalista ja ei-verbaalista modaliteetista, sanojen ohella myös *puheen prosodiset piirteet* voivat olla tärkeitä audiotallenteiden sisällönkuvailussa ja tiedonhaussa. Sisällönkuvailulla tässä yhteydessä tarkoitetaan lähinnä audiotallenteiden *indeksointia* puheen prosodisten piirteiden avulla kuten esim. puhujan sukupuoli (mies tai nainen) tai ikä karkealla asteikolla ilmaistuna kuten esim. onko puhuja lapsi tai aikuinen. Yleensä ottaen puheen prosodisilla piirteillä tarkoitetaan suprasegementaalisia eli sanantasolle ja sen yli lokalisoituvia puheen ominaisuuksia ja niiden mitattavia akustisia korrelaatteja kuten esim. ilmauksen sävelkulkua (intonaatiota), duraatiota

*Kirjoittajat työskentelevät Oulun yliopistossa MediaTeam-tutkimusryhmässä sen Language and Audiology Team:ssä. Lisätietoja osoitteesta http://www.mediateam.oulu.fi?/lang=fi

(kestoa), painoa (emfaasia) ja tauotusta. Puheen prosodisten piirteiden avulla voidaan jopa kääntää ilmauksen merkitys täsmälleen päin vastaiseksi: puhuja saattaa esim. sanoa ”Kyllä”, mutta ilmauksen todellinen merkitys saattaakin olla ”Ei”. Tällöin vaikka meillä olisi täydellisesti toimiva puheentunnistusjärjestelmä, ilman puheen prosodisten piirteiden analyysiä, emme välttämättä saisi vielä tulkittua kaikkea audiotallenteiden todellista merkitystä sisältöanalyysissä tai puhetiedonhaussa.

Puheen prosodisia piirteitä voidaan hyödyntää sekä audiotallenteiden sisällönkuvailussa että prosodisten piirteiden avulla tapahtuvassa tiedonhaussa (litteroitujen) sanojen ohella tai mahdollisesti (osittain) jopa niiden asemesta. Tämä on mahdollista, koska puheen prosodiset piirteet välittävät runsaasti tietoa sekä puhujasta itsestään, esim. puhujan koulutustaustasta tai iästä, että tietoa puhetilanteesta, esim. siitä onko puhetilanne muodollinen vai epämuodollinen (Laver 1994, 14). Kaikki tämä tieto on potentiaalisesti tärkeää audiotallenteiden indeksoinnissa ja puhetiedonhaussa.

Sisällönkuvailun kannalta puheen prosodiset piirteet voivat identifioida sellaisia puhujaominaisuuksia kuten esim. puhujan ikä, sukupuoli, sekä diskurssin globaaleja tai lokaaleja ominaisuuksia kuten esim. painotusta sisältäviä diskurssin kohtia. Niiden potentiaalisesta hyödyllisyydestä huolimatta puheen prosodisia piirteitä ei juurikaan hyödynnetä vielä audiotallenteiden sisällön kuvailussa prosodisten indeksitermien avulla tai puheen prosodisiin piirteisiin perustuvassa tiedonhaussa (Garofolo et al. 2000, 16). Vielä toistaiseksi audiotallenteiden sisällönkuvailu näyttää pääosin keskittyvän puheen ja muun (ympäristö)äänen erottamiseen toisistaan tai puheen ja musiikin erottamiseen toistaan (ks. Penttilä et al. 2001), mikä sinällään voi olla myös tarpeen eri sovellusalueilla.

Tämän artikkelin tarkoitus on yrittää kartoittaa, millaisia sisältöjä voidaan tunnistaa puheen prosodisten piirteiden avulla audiotallenteista ja miten puheen prosodisia piirteitä voitaisiin hyödyntää audiotallenteiden sisällön kuvailussa ja tiedonhaussa. Puhetiedonhaun nykytilaa käsitellään myös. Metodologisia kysymyksiä kuten esim. sisällönkuvailun ja tiedonhaun kannalta tärkeiden prosodisten piirteiden akustisia korrelaatteja sekä identifioitujen sisältöjen yksiselitteisyyttä/monitulkintaisuutta käsitellään myös.

Puhetallenteiden sisällönkuvailun ja tiedonhaun lisäksi puheen prosodiset piirteet ovat tärkeitä myös *ihmisen ja tietokoneen välisessä vuorovaikutuksessa*: tietokoneiden inhimillistämisessä ei vielä riitä, että tietokone pelkästään vastaa selvällä ja kuuluvalla äänellä, vaan sen tulisi kyetä myös osoittamaan emootioita ja asenteita sekä olla kärsivällinen vuorovaikutuksessa ihmisen kanssa. Puheen prosodiset piirteet voivat olla tärkeässä roolissa myös tekniikan inhimillistämisessä ihmisen ja tietokoneen välisessä vuorovaikutuksessa. (Morton 1996.)

2. Mitä puheen prosodiset piirteet voivat kertoa meille audiotallenteiden sisällöstä?

2.1. Puhujaominaisuuksien tunnistus audiotallenteesta

Puheen prosodisten piirteiden analyysin avulla voidaan identifioida *puhujaominaisuuksia* kuten esim. puhujan ikä ja sukupuoli. Puhujan *ikä* voidaan tunnistaa karkealla asteikolla siten, että puhuja on joko lapsi, aikuinen tai vanhus. Balázs:n (1994, 86-88) mukaan puheäänen akustisten tutkimusten perusteella on voitu todeta, että äänen spektraaliset ominaisuudet muuttuvat huomattavasti iän myötä, erityisesti 60:n ikävuoden jälkeen. Esimerkiksi iäkkäiden ihmisten puheen perussävel eli F0 eroaa aikuisten puheen perussävelestä siten, että vanhemmilla puhujilla formanttien eli voimistuneiden osasävelalueiden reunat ovat epämääräisempiä kuin nuoremmilla puhujilla. Puhetietokantojen lisäksi iällä katsotaan olevan merkitystä myös tekstiaineistojen luokittelussa: tyypiteltäessä tekstejä tiettyihin typologioihin, kirjoittajan iällä on merkitystä, koska lasten ja vanhusten tuottamat tekstit eroavat lingvistisesti muun aikuisväestön tuottamista teksteistä (Sinclair & Ball 1996, 12). Tällä tarkoitettaneen sitä, että lapsi on esim. 10-vuotias ja aikuinen 60-vuotias, vaikka se ei tarkkaan ottaen käykään ilmi alkuperäisestä lähteessä.

Puheen prosodisten piirteiden analyysillä puhujan iän määrittämisessä saattaa olla myös *oikeudellista merkitystä* kuten kävi ilmi Eeva Vuoren tapauksessa: Korkeimman oikeuden tuomarille Eeva Vuorelle soitettiin herjaava puhelu, jonka hän nauhoitti. Puhelun soittajaksi epäiltiin

aluksi Turun kaupungin entistä kaupunginjohtajaa Juhani Leppää, joka kuitenkin kiisti soittaneensa kyseisen puhelun. Tällöin Turun yliopiston Fonetiikan laitos analysoi kyseisen nauhoitteen ja kykeni osoittamaan, että puhelun soittaja oli vanhempi mies kuin Juhani Leppä; myöhemmin soittajaksi paljastui Juhani Lepän iäkäs isä.

Puhujan *sukupuoli* voidaan tunnistaa jopa useiden akustisten korrelaattien avulla. Puheen perussävelen eli *F0:n vaihtelu* on suurempaa naispuhujilla kuin miespuhujilla (Johns-Lewis 1986, 212; Graddol 1986). Miesten ja naisten puheen (keskimääräinen) F0 eroaa myös huomattavasti toisistaan: miespuhujien F0 vaihtelee tyypillisesti 60-240 Hz:n välillä. Keskimääräinen puheen perussävel miespuhujilla on 120 Hz, naisilla se on 225 Hz (Cruttenden 1986, 4). Puhujan sukupuoli voidaan Carey et al.:n (1996, 1800) mukaan tunnistaa pelkästään puheen perussävelen parametrien avulla 98%:sti oikein kielestä riippumatta.

2.2. Diskurssin rakenteen globaalisten ja lokaalisten piirteiden identifiointi puheen prosodisten piirteiden avulla

Puhujaominaisuuksien lisäksi puheen prosodisten piirteiden avulla voidaan myös tunnistaa diskurssin rakenteen *globaaleja* ja *lokaaleja* piirteitä. Selkeä diskurssin globaalinen rakenne, esim. puheen paratoonirakenteen muodossa, jolla rakennetaan tekstiä prosodisin keinoin, luonnehtii joitakin diskurssityyppejä kuten esim. uutislähetykset (Iivonen et al. 1987, 246) tai identifioi joitakin *puhetilannepiirteitä* kuten esim. onko puhe etukäteen (huolellisesti) valmisteltua tai spontaania puhetta. Puheen *paratoonirakenteella* tarkoitetaan puhekielen paragrafeja, jotka ovat analogisia kirjoitetun kielen kappalerakenteelle (Brown & Yule 1983, 101). Puhetilannepiirteitä kuten esim. sitä, onko puhe etukäteen huolellisesti valmisteltua vai spontaania puhetta, voidaan identifioida/mitata *puheen ja hiljaisuuden välisellä suhteella*, joka vaihtelee puheen spontaanisuusasteen mukaan siten, että etukäteen valmistellussa puheessa kuten esim. poliittinen puhe, puheen ja hiljaisuuden välinen suhde on pienempi kuin spontaanissa puheessa, jossa esiintyy enemmän taukoja (Johns-Lewis 1986, 204.)

Puheen ja hiljaisuuden välistä suhdetta voidaan hyödyntää myös sovelluksissa, joissa on tarpeen arvioida sisältöjen vastaavuutta kun materiaali on käännetty kielestä toiseen *simultaanitulkkauksen* avulla. Mittaamalla sitä, missä määrin simultaanikäännös jää jälkeen alkuperäistä puheesta puheen ja hiljaisuuden välisen suhteen erilaisina parametreina voidaan tehdä päätelmiä siitä, missä määrin alkuperäisen ja simultaanitulkatun puheen sisältö vastaavat toisiaan (Yagi 1999). Jos tulkkaus jää ajallisesti paljon jälkeen tulkattavasta puheesta, sen sisältö ei ehkä vastaa tarkalleen alkuperäistä (tosin täysin eksakteja käännöksiä ei liene olemassakaan). Tällöin esim. TV-toimittaja voisi arvioida haastattelumateriaalin luotettavuutta sen perusteella, missä määrin alkuperäinen ja tulkattu versio eroavat toisistaan ajallisesti.

Puhenopeuden vaihtelu (nopeutus tai hidastus) fraasien, virkkeiden tai kappaleiden sisällä identifioi myös diskurssin lokaalisia piirteitä kuten esim. erilaisia *lausetyyppejä*, esim. alisteisissa lauseissa puhenopeus on suurempi kuin itsenäisissä lauseissa (Crystal 1969, 152-153). Näiden piirteiden paikantamiseen audiotallenteissa tarvittaisiin puhenopeuden mittaria, joka tarvitsee vain kyetä paikantamaan puhenopeuden muutoskohdat.

Painotusta käsittävät puheenkohdat ovat myös potentiaalisesti tärkeitä prosodisten piirteiden avulla tapahtuvan sisällönkuvailun ja tiedonhaun kannalta. Niiden avulla voidaan esim. laatia *auditorisia yhteenvetoja* tiettyjen diskurssityyppien kuten esim. urheilukilpailujen selostusten huippuhetkistä ilman, että tarvitsisi kuunnella läpi koko puhetallennetta näiden kohtien löytämiseksi. Jollakin tapaa painottuneella puheella on useitakin akustisia korrelaatteja, joiden avulla ne voidaan identifioida suhteellisen helposti. Esim. yhtä viidesosasekuntia lyhyemmät *tauot* esiintyvät painottuneen puheen yhteydessä (Argyle 1975, 355). Puheen *intensiteetti* eroaa myös tyypillisesti selkeästi ympäristöstään puhuttaessa normaalia kovemmalla äännellä kuten esim. urheilukilpailujen selostusten huippuhetkissä. Ympäristöstään selkeästi eroava puheen intensiteetin taso voidaan todeta esim. analysoimalla muutos joltakin intensiteetin normaalitasolta.

Karlgrenin (1961, 674) mukaan puhenopeus korreloi puheen *informatiivisen sisällön* kanssa siten, että ”tyhjä” puhe, joka vaatii vähemmän mentaalista suunnittelua (esim. jargon, slangi tai läpiopitut fraasit), voidaan puhua nopeammin kuin ilmaukset, joiden sisältö vaatii enemmän

mentaalista suunnittelua.

Joitakin epäröinti-ilmiöitä kuten esim. niin sanottua *yneemiä* pidetään asiantuntijan merkkinä, ja sen esiintyminen korreloi puheen suunnittelun tarpeiden kanssa, mikä taas voi korreloida puheen intellektuaalisen sisällön kanssa. Yneemiä saatetaan tavata runsaastikin joidenkin puhujien puheessa kuten esim. presidentti Mauno Koiviston puheessa. Analysoidessaan sanojen frekvenssiä puhutussa ja kirjoitetussa kielessä *British National Corpus*:ssa Leech et al. (2001) toteavat, että soinnilliset täytetyt epäröintitauot, joissa esiintyy vokalisaatioita kuten esim. ”er” ja ”erm” ovat yleisiä ”ajatustaukoja” englannin kielessä huolellisesti etukäteen suunnittelussa julkisessa puheessa. Mallintamalla näitä vokalisaatioita akustisesti ja tunnistamalla ne esim. avainsanoja tunnistavan puheentunnistusjärjestelmän avulla, olisi ehkä mahdollista tunnistaa puhetallenteista jonkinlainen yleinen asiantuntijasisältödimensio sen lisäksi, että voitaisiin erottaa myös se onko puhe etukäteen valmisteltua vai spontaania puhetta.

Toisaalta kuten Hawkins (1973, 243) osoittaa, sujuva puhe, jossa puhenopeus on nopeampi ja jossa esiintyy vähemmän epäröinti-ilmiöitä, ei välttämättä aina korreloi puheen vähäisen informatiivisen tai intellektuaalisen sisällön kanssa; puheen sujuvuus voi yhtä hyvin osoittaa joko puhujan kokeneisuutta esiintyjänä tai suurempaa itseluottamusta tai molempia. On myös mahdollista puhua sujuvasti aiheesta, jonka tuntee läpikotaisin.

Yllä olevat esimerkit valaisevat erästä puheen prosodisten piirteiden avulla tapahtuvaa sisällönkuvailun ja tiedonhaun metodologista perusongelmaa eli tulosten osittaista moni-tulkintaisuutta tai epätarkkuutta, josta enemmän puhetiedonhaun nykytilaa käsittelevän luvun jälkeen.

3. Puhetiedonhaun nykytilasta

Nykyisin *puhetiedonhaussa* hyödynnetään perinteisen tekstihakupohjaisen tiedonhaun menetelmiä puheentunnistuksen apuna (Ng 1996, 20). Kuten johdannossa mainittiin tämän lähestymistavan ongelmana ovat mm. puheentunnistimen tekemät virheet. Tekstihakuun kehitetyt järjestelmät eivät myöskään sellaisenaan sovellu puhetiedonhakuun, koska ne olettavat (implisiittisesti), että puheentunnistuksen avulla tuotettu tekstitranskriptio on virheetön (Ng 1996, 20). Puheentunnistuksen tuottaman tekstitranskription tunnistusvirheitä voidaan jossain määrin vähentää käyttämällä erilaisia virheenkorjaustekniikoita. Puhetiedonhaun keskeisimpiä tutkimuskysymyksiä on nykyisin se, miten puhetallenteiden sisältö pitäisi kuvata tavalla, joka mahdollistaa sekä tehokkaan tiedontallennuksen että tiedonhaun (Ng 1996, 20).

Oleellisilta osiltaan puhetiedonhaku on samanlaista kuin tekstihakukin: dokumentit, olivatpa ne sitten tekstiä tai puhetta, täytyy ensin indeksoida jotenkin, hakukyselyt täytyy muotoilla ja tiedonhaun ytimenä on hakukyselyjen ja dokumenttien tallennusmuodon/indeksien välinen vastaavuus kuten tiedonhaussa tavallisestikin, mikä mahdollistaa tiedonhaun käytännössä (Ng 1996, 20-21).

Jotta vältyttäisiin puheentunnistimen tekemiltä virheiltä sanojen tunnistuksessa, voidaan käyttää myös sanoja pienempiä yksiköitä kuten esim. tavujen kaltaisia yksiköitä (Ng 1996, 22) tai foneemeja (Ferrieux & Peillon 1999) indeksitermeinä sekä puhetallenteiden indeksoinnissa että varsinaisessa puhetiedonhaussa (Ng 1996, 22). Näiden yksiköiden etuna on se, että niiden määrä on huomattavasti pienempi kuin sanojen määrä. Indeksoinnissa voidaan esim. käyttää vokaali-konsonantti-vokaali – piirteitä (VCV-piirteitä), jossa vokaalien välissä esiintyy maksimaalinen määrä konsonantteja, esim. sanalla INFORMATION on seuraavat VCV-piirteet: INFO, ORMA ja ATIO (Ng 1996, 22). VCV-piirteiden heikkoutena, erityisesti jos indeksipiirteet valitaan tekstistä puheesta tunnistetun transkription asemesta, on *akustisesti samankaltaisten* sanojen esiintyminen (puheentunnistuksessa sanaston koon ja akustisesti samankaltaisten sanojen määrän välillä voidaan saavuttaa optimitasapaino, ks. Rosenfeld 1995).

Indeksipiirteet voidaan myös valita akustisen datan perusteella tai vaihtoehtoisesti puheentunnistimen tuottaman tekstitranskriptioon pohjautuen, jolloin akustisesti samankaltaisia sanoja esiintyy vähemmän. Tällöin voidaan myös huomioida puheentunnistimen itsensä ominaisuudet, esim. siinä käytetyn sanaston koko, joka vaikuttaa sen tunnistustarkkuuteen. Erityisten avainsanojen ja sanoja pienempien yksiköiden asemesta voidaan myös käyttää yksittäisiä foneemeja. Tällöin foneemitunnistimen tuottamia foneemisekvenssejä voidaan prosessoida ja erottaa niistä indeksipiirteitä.

Indeksipiirteet tässä lähestymistavassa audiotallenteiden indeksointiin ovat tyypillisesti *foneemisekvenssejä*. Tämän lähestymistavan etuna on se, että käytetty sanasto ei rajaa tunnistustuloksia sanastossa esiintymättömien sanojen muodossa, jolloin tunnistin voi toimia useilla eri käyttöalueilla. Indeksitermien määrää ei myöskään rajoita puheentunnistimen kapasiteetti eli käytössä oleva sanasto ja sen laajuus. Foneemitunnistimen tunnistustuloksia voidaan myös *jälkikäsitiellä* erilaisten tunnistusvirheiden määrän minimoimiseksi. (Ng 1996, 23.)

Millaisia tuloksia on sitten saatu käytännön sovelluksilla puhetiedonhaussa? Sprack Jones et al.:n (2001, 28) mukaan nykyaikaisen puheentunnistimen tuottamasta tekstitranskriptiosta tehty tiedonhaku voidaan tehdä yhtä tehokkaasti kuin varsinaisesta oikein litteroidusta tekstistäkin. Hakutuloksiin eivät välttämättä vaikuta juurikaan puhedatan virheellisyydet tai puheentunnistimesta itsestään riippuvat tekijät kuten tunnistimessa käytetyn sanaston koko. Hakukyselyjen laajennus on myös osoittautunut tehokkaaksi keinoksi parantaa hakutuloksia. Samoilla linjoilla näyttäisivät olevan myös muutkin tutkijat, esim. Garofolo et al. (2000), järjestelmien suorituskyvystä, joissa yhdistetään puheentunnistusta ja tiedonhaun menetelmiä kuten yllä kuvattiin: järjestelmien suorituskyky on usein jo varsin hyvä ja sitä voidaan arvioida kvantitatiivisesti.

Puheenprosodisten piirteiden hyödyntäminen, esim. puhuja- tai puhetilannekohtaisten piirteiden osalta kuten yllä kuvattiin, näyttäisi kuitenkin Garofolo et al.:n (2000, 16) mukaan olevan vielä lähes hyödyntämättä. Juuri puheen prosodiset piirteet mahdollistavat nykyisten tiedonhakujärjestelmien laajentamisen uusilla puhetallenteiden sisältöä kuvaavilla indeksitermeillä (Ng 1996, 20).

Tällaista puheen prosodisten piirteiden avulla tapahtuvaa audiotallenteiden indeksointia voisi käyttää luultavasti myös audiotallenteiden karkeassa esiluokittelussa tiedonhakua varten. Seuraavaksi käsitellään metodologisia ongelmia puheen prosodisten piirteiden hyödyntämisessä tiedonhaussa.

4. Metodologisia ongelmia puheen prosodisiin piirteisiin perustuvissa audiotallenteiden sisällönkuvailuissa

Kuten yllä jo todettiin puheen prosodisten piirteiden avulla tapahtuvassa sisällön kuvailussa ja tiedonhaussa on joitakin metodologisia ongelmia kuten esim. joidenkin prosodisten piirteiden akustisten korrelaattien mallinnus sisällön kuvailun tarpeisiin ja tunnistettujen sisältöjen osittainen monitulkintaisuus.

Lähtökohtaisesti puheen prosodiset piirteet vaihtelevat *suprasegmentaalisesti* eli pitemmällä aikaikkunalla kuin äännetason akustiset ilmiöt. Käytännössä tämä voi tarkoittaa sitä että, puheen prosodisten piirteiden avulla tapahtuvaa sisällönkuvailua ja tiedonhakua varten voidaan joutua kehittämään omia analysointitekniikoita, joiden implementaatio ei sinällään ole välttämättä vaikeaa johtuen tekniikoiden karkeasta toimintatavasta kuten esim. puhenopeuden muutoskohtien tunnistamien audiotallenteesta.

Eräs keskeinen puheen prosodisten piirteiden ominaispiirre on niiden avulla identifioitujen sisältöjen/piirteiden *monitulkintaisuus* (Polzin 1999, 5). Esim. puhenopeuden vaihtelu tavujen, sanojen tai virkkeiden sisällä voi heijastella puheensuunnittelun tarpeiden asemesta myös erilaisia *emootioita* kuten esim. kiihtymystä tai kärsimättömyyttä (Crystal 1997, 248); sinällään emootioita sisältävät puheenkohdat saattavat myös olla mielenkiintoisia sisällönkuvailun ja tiedonhaun kannalta. Eräs tapa hallita prosodisten piirteiden aiheuttamaa monitulkintaisuutta on Polzin:n (1999, 9) mukaan mallintaa eksplisiittisesti useita prosodisten piirteiden funktioita samanaikaisesti.

Jotkut puheen prosodiset piirteet tai paranlingvistiset piirteet ovat myös *kielikohtaisia*, jolloin niillä saattaa olla erilainen tulkinta eri kielissä: esim. englannin kielessä sorainen äänenlaatu välittää konnotaatioita kuten esim. välinpitämättömyyttä tai halveksuntaa kun taas suomen kielessä ko. äänenlaatu on yleinen, erityisesti miehillä, ilman ennen mainittuja konnotaatioita (Crystal 1997, 171). Toisaalta jotkut prosodisten piirteiden akustiset korrelaatit kuten esim. sanojen äännerakenteen taipumus redusoitua kohti konsonantti-vokaalirakennetta nopean puheen vaikutuksesta (Hatch 1983, 27), saattavat olla myös kielestä (ja puhujasta)

riippumattomia, jolloin niitä voidaan hyödyntää useissa kielissä.

Puheen prosodisten piirteiden rooli (perinteisessä) tiedonhaussa näyttää myös selvästi muodostuvan entistä keskeisemmäksi. On alettu mieltää, että jopa (vaikeasti mallinnettavat) spontaanin puheen ominaisuudet tarjoavat itse asiassa myös hyödyllistä tietoa esim. puheentunnistusta varten (Juan 1998, 41) puheen prosodisiin piirteisiin perustuvan sisällönkuvailun ja tiedonhaun lisäksi.

Eräs puheen prosodisten piirteiden keskeinen kielestä riippumaton funktio on *ilmausten prosodinen jaksottelu* syntaktisella tasolla, jota voidaan hyödyntää audiotallenteiden jaottelussa lingvistisesti mielekkäisiin kokonaisuuksiin (Shriberg & Stolke 2000, 4). Sinällään analysoitavan aineiston segmentointi edustaakin erästä keskeistä ongelmaa monilla kieliteknologian sovellusalueella.

Se missä järjestyksessä puheen prosodisten piirteiden analyysi itse asiassa tehdään on myös merkityksellinen, koska osa puheen prosodisista rakenteesta selittyy esim. *lingvistisen akkommodaation* kautta: jos puhujina ovat esim. aikuinen ja lapsi, niin aikuisen hidas puhenopeus selittyy sillä, että lapselle puhuttaessa puhutaan yleensä puolet hitaammin kuin aikuiselle puhuttaessa. Tällöin hitaan puhenopeuden muut syyt voidaan sulkea pois (disambiguoida) analysoimalla ensin puhujien sukupuoli ja ikä karkealla asteikolla sekä hyödyntämällä tietoa kielellisen kommunikaation peruspiirteistä.

5. Päätäntä

Huolimatta tulosten osittaisesta monitulkintaisuudesta ajatus puheen prosodisten piirteiden avulla tehtävästä audiotallenteiden automaattisesta sisällönkuvailusta prosodisten indeksitermien avulla on houkutteleva; yksinkertaisimmillaan sen avulla on mahdollista esim. jaotella suuri määrä audiotallenteita lingvistisesti mielekkäisiin osiin, joiden sisällä voidaan sitten tehdä muita tarvittavia analyysejä. Erityisesti normaalista poikkeavat puheen prosodiset piirteet kuten esim. painottuneen puheen akustiset korrelaatit ovat helpoimmin implementoitavia piiteitä sisällönanalyysin tarpeisiin.

Puheen prosodiset piirteet voivat välittää informaatiota sekä puhujista itsestään, esim. heidän tunnetilastaan puhehetkellä, puhetilanteesta, onko se esim. muodollinen vai epämuodollinen, että dokumenttien sisällöstä paikantamalla esim. jollakin tapaa korostunutta informaatiota audiotallenteen sisällä. Puheen prosodisten piirteiden avulla tehtävä tiedonhaku voi olla hyödyllistä esim. silloin kun halutaan löytää audiotallenteen sisältä puhujan jollakin tapaa korostuneesti esiin tuomaa informaatiota tai kun halutaan löytää audiotallenteita, joissa esim. neljä aikuista ihmistä keskustelee keskenään jossakin muodollisessa puhetilanteessa. Puheen prosodisia piirteitä voi myös hyödyntää audiotallenteiden karkeassa luokittelussa puhuja- tai puhetilannepiirteiden perustella, esim. spontaania puhetta käsittäviin audiotallenteisiin tai jollakin tapaa strukturoituihin audiotallenteisiin.

Riippuen siitä, mitä analyysissä on itse asiassa saatu selville, on mahdollista tehdä lisäanalyysejä, esim. jonkin asiantuntijajärjestelmän avulla, hyödyntäen muuta lingvististä tietoa tai kommunikaation peruslainalaisuuksia. Kielestä riippumattomat puheen prosodiset piirteet skaalautuvat myös moniin eri kieliin. Tiedonhaun lisäksi puheen prosodisten piirteiden avulla tehtävästä sisältöanalyysistä voisi olla hyötyä myös puheentunnistusjärjestelmille, joissa tunnistus tehdään useammassa vaiheessa. Ihmisen ja tietokoneen välisessä interaktiossa puheen prosodisia piirteitä voidaan niin ikään hyödyntää puhesynteesissä kun yritetään kehittää tietokoneita, jotka kykenevät osoittamaan erilaisia asenteita ja emootioita.

Hyväksytty julkaistavaksi 15.11.2002

Viitteet

Argyle, M. (1975). *Bodily Communication*. London: Methuen & Co. Ltd.

Balázs, B. (1994). Voice Quality Changes in Old Age. *Acta Linguistica Hungarica* 42(1-2):83-92.

Brown, G. & Yule, G. 1983. *Discourse Analysis*. Cambridge: Cambridge University Press.

Carey, M., Parris, E., Lloyd-Thomas, H., Bennett, S. (1996). Robust Prosodic Features for Speaker Identification. *Proc. ICSLP '96*. Vol. 3:1800-1803.

Colineau, N. & Halber, A. (1999). A Hybrid Approach to Spoken Query Processing in Document Retrieval System. 7th of November 2002. http://svr-www.eng.cam.ac.uk/~ajr/esca99/Colineau.pdf.

Cruttenden, A. (1986). *Intonation*. Cambridge: Cambridge University Press.

Crystal, D. (1969). *Prosodic Systems and Intonation in English*. Cambridge Studies in Linguistics 1, Cambridge, Cambridge University Press.

Crystal, D. (1997). *The Cambridge Encyclopaedia of Language*. Cambridge: Cambride University Press.

Ferrieux, A. & Peillon, S. (1999). Phoneme-Level Indexing for Fast and Vocabulary-Independent Voice/Voice Retrieval. 30th of October 2002. http://svr-www.eng.cam.ac.uk/~ajr/esca99/Ferrieux.pdf.

Graddol, D. (1986). Discourse Specific Pitch Behaviour. *Intonation in Discourse*. Toimittanut Johns-Lewis, C. London & Sidney: Croom Helm.

Garofolo, J. S., Auzanne, C. G. P., Voorhees, E. M. (2000). The TREC Spoken Document Retrieval Track: A Success Story. In *RIAO'2000 Conference Proceedings: Content-Based Multimedia Information Access*.

Hatch, E. M. (1983). *Psycholinguistics: A second Language Perspective*. Rowley, London, Tokyo: Newbury House Publishers, INC.

Hawkins, P. R. (1973). The Influence of Sex, Social Class and Pause-Location in the Hesitation Phenomena of Seven-Year-Old Children. *Class, Codes and Control*. Vol. 2. Toimittanut Bernstein, B. London and Boston: Routledge & Kegan Paul.

Hindus, D., Schmandt, C., Horner, C. (1993). Capturing, Structuring, and Representing Ubiquitous Audio. *ACM-Transactions-on-Information-Systems*. Vol.11. No. 4: 376-400.

Iivonen, A., Nevalainen, T., Aulanko, R., Kaskinen, H. (1987). *Puheen intonaatio*. Helsinki: Gaudeamus.

Johns-Lewis, C. (toim.) (1986). *Intonation in Discourse*. London & Sidney: Croom Helm.

Juang, B. H. (1998). The Past, Present, and Future of Speech Processing. *IEEE Signal Processing Magazine*. Vol. 15. No. 3:24-48.

Karlgren, H. (1961). Speech Rate and Information Theory. *Proceedings of the Fourth International Congress of Phonetic Sciences*. Helsinki: Mouton & Co.

Laver, J. (1994). *Principles of Phonetics*. Cambridge: Cambridge University Press.

Leech, G., Rayson, P., Wilson, A. (2001). *Word Frequencies in Written and Spoken English: based on the British National Corpus*. London: Longman. 4th of March 2002. http://www.comp.lancs.ac.uk/ucrel/bncfreq/

Morton, K. (1996). Speech Output in HCI Technologies. *IEE Colloquium (Digest) n* 126: 3/1-3/3.

Ng, K. 1996. *Survey of Approaches to Information Retrieval of Speech Messages*.

Penttilä, J., Peltola, J., Seppänen, T. (2001). A Speech/Music Discriminator-Based Audio Browser with a Degree of Certainty Measure. *Proc. Infotech Oulu International Workshop on Information Retrieval (IR-2001)*. Infotech Oulu: 125-131.

Polzin, T. S. (1999). *Detecting Verbal and Non-Verbal Cues in the Communication of Emotions*. Unpublished Ph.D. thesis. School of Computer Science. Carnegie Mellow University.

Rosenfeld, R. (1995). Optimizing Lexical and Ngram Coverage via Judicious Use of Linguistic Data. *Proc. Eurospeech'95*.

Shriberg, E., Stolcke, A. 2001. Prosody Modelling for Automatic Speech Understanding: An Overview of Recent Research at SRI". 16th of August 2002. www.speech.sri.com/papers/prosody2001-overview.ps.gz.

Sinclair, J. McH. & Ball, J. (1996). *EAGLES Preliminary Recommendations on Text Typology*. EAGLES Document EAG-TCWG-TTYP/P. Version of Jun, 1996.

Spräck Jones, K. Jourlin, P., Johnson, S. E., Woodland, P. C. 2001. *The Cambridge Multimedia Document Retrieval (MDR) Project: Summary of Experiments*. Technical Report 517.

Yagi, S. M. (1999). Computational Discourse Analysis for Interpretation. *META* XLIV 2. 4th of June 2001. http://www.erudit.org/erudit/meta/v44n02/yagi/yagi.htm.