KATSAUS

*Erkka Leppänen*

# PUHEDOKUMENTTIEN INDEKSOINTI JA HAKU

Leppänen, Erkka. Puhedokumenttien indeksointi ja haku. [Speech indexing and retrieval. Informaatiotutkimus 19 (3), s. 90-97, 2000.

This article makes a general overview to speech retrieval. Speech retrieval covers both retrieval of spoken documents by spoken or text-based queries and retrieval of any documents by spoken queries. The development of speech retrieval methody is necessary because the amount of spoken data is growing all the time, and speech is richer and more expressive than text. Theoretical readiness for speech retrieval already exists. Automatic speech recognition can be accomplished by identifying either individual words or phonemes. Indexing and retrieval of spoken documents can use the same methods as in text retrieval, but speech retrieval also has some special problems which still hasn't been solved. The article also briefly introduces research made in spoken document retrieval and the special challenges of speech retrieval in Finnish language.

*Address: Erkka Leppänen, University of Tampere, Department of Information Studies, FIN-33014 University of Tampere, Email sterle@uta.fi*

## 1. Johdanto

**Puhehaku** (*speech retrieval*) voidaan määritellä kahdella eri tavalla. Se voi tarkoittaa mitä tahansa tiedonhakua, joka tehdään puhekäyttöliittymän avulla, kohdistuipa haku puheeseen, tekstiin, kuviin, multimediadokumentteihin tai johonkin muuhun aineistoon. Se voi tarkoittaa myös millaista tahansa tiedonhakua, joka kohdistuu puhetta sisältäviin dokumentteihin. (Oard 2000). Tämän artikkelin yhteydessä puhehaku määritellään sanan laajimmassa merkityksessä kattamaan sekä puhekäyttöliittymän avulla tehdyt haut että puhetta sisältävien dokumenttien haut, toisin sanoen kaikki haut, joissa hakuprosessiin kuuluu jossain vaiheessa automaattinen puheenkäsittely.

Interaktiivinen tiedonhaku, jossa tietokone viestii puhumalla tai vastaanottaa puhemuotoisia komentoja tai tekee molempia, on tämän määritelmän mukaan puhehakua. Tällaisessa haussa hakukone joko tulkitsee puhuttua kieltä tai muuntaa kirjoitetun tekstin puheeksi. Kehittynyt käyttöliittymä voi jopa keskustella tiedonhakijan kanssa ja antaa relevanssipalautetta puhemuodossa. Haettavien dokumenttien mediamuodolla ei ole tällöin merkitystä. (Wang 1998)

Puhehakua on myös puhetta sisältävien dokumenttien haku tietokannasta. **Puhedokumentteja** (*spoken document*) ovat esimerkiksi radiolähetykset, tv-ohjelmat ja muut videotallenteet, nauhoitetut luennot, äänikirjeet yms. Puhe voi olla yhtä hyvin monologia kuin kahden tai useamman ihmisen keskustelua. Kysymys on puhehausta, vaikka hakukysymys syötettäisiin tekstimuodossa eikä puhumalla. Myös multimediadokumenttien hakuun liittyy usein puhehakua. (Wang 1998) Multimediadokumentit sisältävät muiden mediatyyppien (teksti, kuva, video) lisäksi yleensä myös auditiivista materiaalia, mikä kattaa sekä musiikin, taustaäänet, äänitehosteet että puheen. Tässä artikkelissa ei kuitenkaan puututa muiden äänielementtien kuin puheen hakuun.

Puhedokumentit tallennetaan nykyään usein sellaisenaan digitoituna audiosignaalina ilman merkittävää tekstimuotoista sisällönkuvailua. Siksi tarvitaan myös tehokkaita automaattisia menetelmiä puhedokumenttien sisällönkuvailuun ja hakemiseen tietokannasta. Puhedokumenttien sisältöperusteisen haun mahdollistava sovellus hyödyttäisi esimerkiksi radion uutistoimitusta, jonka pitää löytää arkistostaan vanhoja lähetyksiään. Puhekäyttöliittymä taas lisäisi huomattavasti tietokantojen käyttötapoja ja sovellusmahdollisuuksia. Puhe on ihmisen luonnollinen tapa viestiä, joten olisi luontevaa saada tietokoneetkin ymmärtämään ja tuottamaan puhetta. Tiedonhakijan on kätevää esittää kysymyksensä puhuen ja vastaanottaa vastauksetkin puheena monissa käytännön tilanteissa, esimerkiksi autoa ajaessa ja muissa tehtävissä, joissa katse ja kädet tarvitaan muuhun toimintaan, vaikka kohteena olisivat teksti- tai kuvadokumentit. Puhuva käyttöliittymä parantaisi myös näkövammaisten ja liikuntarajoitteisten mahdollisuuksia käyttää tietokonetta.

Toistaiseksi sujuvasti puhuvat ja kuuntelevat tietokoneet ovat tuttuja ainoastaan television science fiction -sarjoista, mutta eri tahoilla jo ponnistellaan näiden visioiden toteuttamiseksi myös todellisessa maailmassa.

## 2. Puhedatan tallentaminen

Puhemuotoista dataa tuotetaan paljon. Erilaiset multimediasovellukset kuten myös videotallenteet sisältävät usein puhetta muiden mediamuotojen ohessa. Jotkut dokumentit, kuten radio-ohjelmat, koostuvat pelkästään puheesta. Sähköpostikin voidaan lähettää ääni- tai videomuodossa.

Puhedatan varastoiminen on kalliimpaa ja hankalampaa kuin tekstin, sillä se vaatii runsaasti tilaa ja tehoja tietokoneelta. Tietokoneiden kasvaneen suorituskyvyn myötä puheen muokkaaminen digitaalimuotoon ja tallentaminen eivät kuitenkaan enää vaadi kohtuuttomasti resursseja .

Puheen tallentamisella tiedonhaun tarkoituksiin on sekä etunsa että haittansa tekstiin verrattuna. Vaikka puheen tallentaminen sellaisenaan ei olekaan helppoa, on se kuitenkin yksinkertaisempaa kuin puheen muokkaaminen tekstiksi. Puhe on ilmaisultaan rikkaampaa, monipuolisempaa ja informatiivisempaa kuin pelkkä teksti, sillä se sisältää puhujan äänen, painotukset sekä äänensävyt ja siitä yleensä pystyy päättelemään myös puhujan mielialan ja sukupuolen. Lisäksi puhedataa on helpompi siirtää esim. puhelinlinjoja pitkin. Toisaalta puhehaussa on kyettävä selviytymään puheen tulkinnassa sattuneista virheistä, sanojen erilaisesta ääntämyksestä ym. ongelmista, jotka ovat vielä suurelta osin ratkaisematta. (Wang 1998)

Koska puhe on helpointa tallentaa alkuperäisessä muodossaan, tarvitaan menetelmiä, joilla se voidaan myös hakea samassa muodossa. Tämä on kuitenkin ongelmallista, sillä puhedataa ei pystytä hakemaan samanlaisin sisältöperusteisin menetelmin kuin tekstiä. Myös puhedokumenttien sisällön manuaalinen tarkastaminen on työlästä, koska tällöin tiedonhakijan täytyisi kuunnella koko dokumentti läpi saadakseen tiedot sen sisällöstä (Jones et al. 1997, 192).

## 3. Puheen automaattinen tunnistaminen

**Puheentunnistus** (*speech recognition*) tarkoittaa prosessia, jossa auditiivinen signaali tulkitaan tekstiksi (Cole et al. 1996). Puheentunnistus on puhehaun onnistumisen ehdoton edellytys - määriteltiinpä puhehaku miten tahansa. Puhetta ymmärtävän tiedonhakujärjestelmän toiminta edellyttää, että käyttöliittymä pystyy tulkitsemaan, mitä tiedonhakija siltä kysyy. Silloinkin kun hakukysymys syötetään tekstimuodossa, dokumentin sisältämästä puheesta on tunnistettava merkkijonot, jotta puhedokumentti pystyttäisiin indeksoimaan sisältöperusteisesti. Puhedokumentin transkription manuaalinen kirjoittaminen on useimmiten taloudellisesti mahdoton urakka käytännön sovelluksissa, joten tulkinnan on tapahduttava automaattisesti.

Puheentunnistus kehittyi pitkin harppauksin 90-luvulla. Vuosikymmenen alussa puheentunnistus oli epäluotettavaa ja edellytti tehokasta laitetta, rajoittunutta sanavarastoa, diskreettiä puhetta, puhujakohtaisesti opetettavaa järjestelmää tai näitä kaikkia. Nykyään ainakin erikoissovelluksiin on mahdollista kehittää hyvin luotettavasti ja luontevasti toimivia puheentunnistusjärjestelmiä. Myös puhesynteesi on parantunut laadultaan jatkuvasti. (Räihä 1999, 18-19)

Nykyisin lähes kaikki puheentunnistus perustuu kätkettyjen Markovin mallien (*hidden Markov models*) käyttöön. Kätketty Markovin malli on tilastollinen malli, joka pystyy valitsemaan kuulemastaan ääneyksiköstä kuten sanasta tai foneemista todennäköisimmän tulkinnan. Malli voidaan opettaa tunnistamaan yhden tai useamman puhujan puhe. Kätkettyjen Markovin mallien kehittäminen on edistänyt puheentunnistusta huomattavasti viime vuosina. (Jones et al. 1996, 32)

Puheen automaattisessa tunnistuksessa on kuitenkin yhä monia ratkaisemattomia ongelmia. Puheentunnistus on luonnostaan altis virhetulkinnoille.

Puheentunnistuksen hankaluuksien perussyy on, että digitaalimuotoon muokattunakaan puhe ei perustu yksiselitteisiin merkkijonoihin kuten teksti. Sama sanakin voidaan lausua usealla eri tavalla puhujasta ja tilanteesta riippuen.

Useat puheentunnistusjärjestelmät perustuvat tiettyyn sanastoon, joka järjestelmä on ohjelmoitu tunnistamaan puheesta ja muuntamaan tekstiksi. Tunnistamisprosessissa ohjelma vertaa puheessa kuulemiaan sanoja sanastonsa sanoihin. Tällöin vain sanastossa olevat sanat pystytään tunnistamaan. Sanastoon kuulumaton sana usein tulkitaan joksikin sanaston sanaksi, jos sanat muistuttavat foneettisesti toisiaan. Lyhyet sanat ovat pitempiä taipuvaisempia virhetulkinnoille. Sitä paitsi puheentunnistusohjelmat pystyvät tunnistamaan vain rajallisen määrän sanoja. Puheen selkeyteen vaikuttaa selkeästi myös se, millaisesta tilanteesta puhe on tallennettu. Formaalista puheesta (esimerkiksi uutislähetyksestä) sanat pystytään tunnistamaan huomattavasti tarkemmin kuin tavallisesta spontaanista keskustelusta. (Jones et al. 1997, 194-195)

Puhutun kielen täydellinen ymmärtäminen vaatisi tietokoneelta sekä puheentunnistuksen hallitsemista että luonnollisen kielen ymmärtämistä. Erillisten yksittäin lausuttujen sanojen tunnistaminen on suhteellisen helppoa, mutta koneen on vaikeampi saada selvää jatkuvasta puheesta. Tiedonhaku asettaa puheentunnistukselle vielä lisähaasteita. Tietokoneen pitäisi pystyä tunnistamaan paitsi yksittäiset sanat myös fraasit, jotta monisanaisetkin avainsanat pystyttäisiin indeksoimaan. (Tämä ongelma on yhteinen tekstihaun kanssa.) Ohjelman pitäisi myös pystyä erottamaan indeksoitava puhe taustahälinästä ja erottamaan eri puhujat toisistaan. Puheen merkityssisällön ymmärtäminen vaatii, että tietokone pystyy jäsentämään lauseetkin, sillä muutoin tietokone ei ymmärrä esimerkiksi ääneen lausuttua hakukysymystä. Puheen kieliopillinen jäsentäminen on kuitenkin tuntuvasti vaikeampaa kuin kirjoitetun tekstin, sillä puhekieli ei noudata kielioppisääntöjä senkään vertaa kuin teksti. Myös sanojen väliset rajat on vaikea havaita puheesta (Schäuble 1997, 126). Hankaluuksia aiheuttavat myös puheessa mahdollisesti esiintyvät eri kielet sekä suunnaton sanamäärä, joka ohjelman pitäisi tunnistaa (Mani et al. 1997, 243).

Jotta päästäisiin eroon sanaston aiheuttamista rajoituksista, on kehitetty järjestelmiä, jotka tunnistavat puheesta sanojen sijasta foneemeja. Foneemi on puhutun kielen pienin merkityksiä erottava funktionaalinen yksikkö. Lyhyimmillään se vastaa vain yhtä kirjainta, mutta se voi olla olla myös useammalla kirjaimella kirjoitettava äänne. Puheentunnistuksessa yleensä pyritään tunnistamaan useamman foneemin sarjoja, joita kutsutaan N-grammeiksi. Muuttuja N ilmaisee tunnistettavien yksiköiden lukumäärän, esimerkiksi bigrammi sisältää kaksi foneemia, trigrammi kolme jne.

Foneemeja tunnistava järjestelmä tuottaa puheen äänneasua vastaavan tekstijäljennöksen kiinnittämättä huomiota siihen, mitä sanoja puheessa esiintyy. Näin syntynyttä dataa voidaan indeksoida suoraan tai käsitellä tekstihaun keinoin, mutta tällä menetelmällä ei päästä yhtä tarkkaan tunnistukseen kuin sanojen tunnistamisella. (Abberley et al. 1999) Foneemeja tunnistavien järjestelmien menestys on ollut varsin vaihteleva, mutta keskimäärin niiden tulokset ovat toistaiseksi olleet huonompia kuin sanoja tunnistavien järjestelmien. Kuitenkin yhdistelemällä samassa järjestelmässä sekä sanojen että foneemien tunnistamista on saavutettu yleensä selvästi parempia tuloksia kuin käyttämällä vain toista menetelmää.

Puheentunnistus on teknisesti helpompaa, jos puhujan henkilöllisyys on selvillä ja jos tunnistusohjelma on etukäteen ohjelmoitu tunnistamaan kyseisen puhujan puhetyyli. (Oard 1998) Mobiileissa ympäristöissä, kuten esimerkiksi matkapuhelimessa, liittymä voidaankin virittää käyttäjälleen. Käytännössä on kuitenkin mahdotonta olettaa, että ohjelma aina tuntisi puhujan etukäteen ja osaisi mukautua tämän ääntämykseen.

Puheentunnistuksen tutkimuksessa tunnistuksen onnistuminen ilmaistaan yleisesti käsitteellä **sanavirhetaso** (*word error rate*). Sanavirhetaso on prosenttiluku, joka kertoo, kuinka suuri osuus puheentunnistusohjelman läpikäymistä sanoista tulkitaan virheellisesti tai jää tunnistamatta. Mitä alhaisempi sanavirhetaso on, sitä paremmin puheentunnistuksessa on onnistuttu. (Cole et al. 1996)

## 4. Puhedokumentin indeksointi

**Indeksoinnilla** tarkoitetaan dokumentin sisältämän datan järjestämistä sellaiseen muotoon, että dokumenttitietuetta voidaan hakea tietokannasta.

Yksinkertaisimmin puhedokumentti voidaan indeksoida teosperusteisesti. Teosperusteisessa indeksoinnissa puhedokumentista kuvaillaan ominaisuuksia, jotka koskevat koko dokumenttia. Viite sisältää tällöin lähinnä tallenteen bibliografiset tiedot (otsikko, puhuja(t), pituus, päivämäärä yms.) sekä luokituskoodit. Dokumentin sisällöstä kerrotaan korkeintaan abstraktilla tai avainsanoilla. Bibliografiset viitteet ovat hallittavissa varsin tavanomaisin menetelmin, eikä tällaisten tietojen luettelointi puhedokumentista poikkea

oleellisesti kirjojen ja muun perinteisemmän aineiston luetteloinnista. Tällöin ei liioin ole tarvetta automaattiselle puheentunnistukselle. Teosperusteinen indeksointi on hyödyllistä ja välttämätöntä, mutta riittämätöntä silloin, kun hakijaa kiinnostavat myös puhedokumentin sisällölliset yksityiskohdat tai kun teosperusteiset attribuutit ovat tiedonhakijalle tuntemattomia etukäteen.

Jos puhedokumenttia halutaan hakea sen sisällön perusteella, se on myös indeksoitava sisältöperusteisesti. Käytännössä tämä tarkoittaa puhedokumentin sisältämien sanojen viemistä käänteistiedostoon. Jos indeksoinnissa käytetään puheentunnistusta, Jonesin et al. (1997, 193) mukaan ainoa kätevä tapa indeksoida uuden puhedokumentin sisältö on suorittaa indeksointi automaattisesti samalla kun dokumentti viedään tietokantaan.

Puhedokumenttien sisältöperusteisessa indeksoinnissa törmätään samoihin ongelmiin kuin tekstidokumenttienkin indeksoinnissa, mutta siinä kohdataan myös koko joukko uusia haasteita. Ilmeisimpiä näistä on se, että puhedokumenttien sisältö on etukäteen tuntematon, ja siksi ensimmäisessä vaiheessa on aina suoritettava indeksointioperaatio automaattista puheentunnistusta käyttäen. Jos puheentunnistus perustuu yksittäisten sanojen tunnistamiseen, operaatio tehdään joko pyrkien tunnistamaan kaikki dokumentissa esiintyvät sanat tai sitten tietty ennakkoon valittu avainsanajoukko. Kummassakin tapauksessa avainsanat rajoittuvat siihen sanastoon, joka puheentunnistusoperaatio on etukäteen ohjelmoitu tunnistamaan. Monet uudet sanat, erityisesti uudet erisnimet, saattavat jäädä tunnistamatta kokonaan, koska ne eivät kuulu tunnistusohjelman sanastoon. Kuitenkin erisnimet ovat tiedonhaun kannalta tärkeitä, sillä niitä käytetään usein hakutermeinä. Tekstitietokannoissa tämä ongelma on helposti ratkaistavissa lisäämällä uuden dokumentin sanat käänteistiedostoon, mutta puhetietokannoissa uuden sanan lisääminen ei ole näin yksinkertaista, sillä tietokantaohjelmiston pitäisi oppia tunnistamaan sanan kirjoitusasun lisäksi myös sen äänneasut. (Schäuble 1997, 126) (Jones et al. 1997, 194)

Avoin kysymys on, onko kannattavampaa indeksoida puheesta kaikki sen sisältämät sanat vaiko vain joukko keskeisiä avainsanoja. Jos puhedokumentin kaikki sanat muokataan tekstiksi, on lopputuloksena tekstidokumentti, jota voidaan hakea tavanomaisilla tekstihaun menetelmillä. Menetelmä kuitenkin edellyttäisi, että tietokone kykenee tunnistamaan valtavan suuren sanamäärän. Sen täytyisi myös suoriutua sanojen erilaisesta ääntämyksestä, eri puhujien äänistä, taustahälystä ym. ongelmista, joista ei tarvitse piitata tekstihaussa. (Wang 1998) Toistaiseksi virheetön puheentunnistus ei ole toteutettavissa.

Jos puhedokumenteista sen sijaan pyritään tunnistamaan vain suppea, etukäteen määritelty joukko avainsanoja, ei tietokoneen tarvitse tuntea niin laajaa sanastoa. Tiettyjen sanojen poimiminen puheesta on huomattavasti helpompaa kuin koko puhedokumentin muokkaaminen tekstiksi. Tällöin kuitenkin pitäisi jo indeksoitaessa tietää, millaista sanastoa dokumentit sisältävät ja millaisia kyselyitä tietokantaan oletettavasti tehdään (Wang 1998). Näin ollen avainsanapoiminta sopii vain aihealueeltaan rajattuihin tietokantoihin. Tämän menetelmän puutteena on myös se, että sanastoon on hankalaa lisätä uusia avainsanoja, kun siihen tulee tarvetta. Yksittäisten avainsanojen indeksointi on tarpeellinen välivaihe puhehaun kehittämisessä, mutta rajoittunut sanavarasto tuskin riittää tulevaisuuden puhehakujen tarpeisiin.

Mahdollisesti tulevaisuudessa puheentunnistuksessa kannattaa keskittyä tunnistamaan foneemeja yksittäisten sanojen sijasta. Foneemitunnistuksessa tunnistetaan puheen äänneasut, ja niiden perusteella puhe tulkitaan ja muokataan sanoiksi. Foneemien tunnistamisessa ei tarvitse rajoittua tiettyyn sanavarastoon, se ei kuormita järjestelmää yhtä paljon, ja se soveltuu sekä puhedatan tallentamiseen että puhemuotoisen hakukysymyksen tulkintaan. Foneemien tunnistuksessa ei vielä nykyisin kuitenkaan päästä yhtä hyvään tarkkuuteen kuin sanojen tunnistuksessa. (Ferrieux & Peillon 1999)

Glavitsch ja Schäuble (1992, 169) esittelevät keskeisiä vaatimuksia puhedokumentista indeksoitavalle datalle:

1. Indeksoitavien ominaisuuksien täytyy olla foneettisia yksiköitä, jotka on helppo tunnistaa puheentunnistusprosessissa.

2. Erilaisten indeksoitavien ominaisuuksien lukumäärän täytyy olla pieni.

3. Indeksoitavien ominaisuuksien tulee olla erottelukykyisiä, ts. niiden perusteella pitää pystyä erottamaan dokumentit toinen toisistaan.

4. Indeksoitavien ominaisuuksien frekvenssi tiedonhaussa ei saa olla liian pieni, jotta saanti pysyy hyvänä ja koska puhehakujärjestelmä tarvitsee runsaasti harjoitusmateriaalia.

Mahdollisia indeksoinnin kohteita ovat foneemit, sanat ja fraasit. Sanat täyttävät vaatimuksen **1.**, koska ne ovat tunnistettavia yksiköitä puhedokumenteissa.

Vaatimus **2.** ei kuitenkaan täyty, sillä puhedokumentit sisältävät yleensä liikaa sanoja, jotta ne kaikki voitaisiin indeksoida. Myös vaatimukset **3.** ja **4.** täyttyvät vain osittain, sillä on olemassa yleisiä sanoja, jotka eivät erottele dokumentteja tarpeeksi hyvin, ja harvinaisia sanoja, jotka eivät esiinny tarpeeksi usein, jotta ohjelma voisi harjoitella niiden tunnistusta. (Glavitsch & Schäuble 1999, 169)

Toisaalta foneemit ovat liian pieni yksikkö indeksoitavaksi. Vaatimus **1.** ei täyty, koska foneemien ääntämys vaihtelee, ja siten niiden tunnistaminen on vaikeaa. Myös foneemien erottelukyky (vaatimus **3.**) on huono. Sen sijaan vaatimukset **2.** ja **4.** täyttyvät, sillä foneemien määrä on suhteellisen pieni ja ne keräävät paljon dokumentteja. Glavitschin ja Schäublen (1999, 169) mukaan paras indeksoitava ominaisuus olisi siten ehkä jossain sanojen ja foneemien välimaastossa.

Jos puhedokumentin sisältö halutaan indeksoida perusteellisesti, on siitä indeksoitava muutakin kuin siinä esiintyvät sanat. Yleensä puhedokumentti sisältää muutakin kuin puhetta: musiikkia, taustaääniä, ja jos kyseessä on video, myös kuvia ja tekstiä. Indeksoinnin kohteita olisivat mm. kohdat, joissa puhuja vaihtuu, taustamusiikki alkaa tai päättyy tai taustalta kuuluu tietty ääni.

Puhedokumenttien indeksoinnin ydinkysymykset ovat siis, kuinka puhedokumentit voidaan indeksoida automaattisesti siten, että tunnistusvirheitä sattuu mahdollisimman vähän, ja kuinka järjestää puhedokumenttien sisältämä data siten että, hakukysymykset voidaan suorittaa tehokkaasti (Schäuble 1997, 122).

## 5. Puhetta tulkitseva käyttöliittymä

Nykyiset käyttöliittymät perustuvat joko tekstiin, grafiikkaan tai molempiin. Seuraava edistysaskel ihmisen ja tietokoneen välisen vuorovaikutuksen kehittämisessä vaatisi kokonaan uudenlaisten vuorovaikutustapojen käyttöönottoa. Käytännössä tämä tarkoittaa lähinnä auditiivisen kanavan eli puheen ja kuulon hyödyntämistä tietokoneiden käyttöliittymissä. (Räihä 1999, 18)

Puhekäyttöliittymä sopisi hyvin tiedonhakuohjelmiston käyttöliittymäksi, koska hakukysymyksen esittäminen puhumalla olisi huomattavasti nopeampaa ja sujuvampaa kuin kirjoittamalla. Puheentunnistusteknologia on jo edennyt niin pitkälle, että perusteet puhekäyttöliittymille ovat jo olemassa. Kuitenkin puheentunnistuksessa on vielä puutteita ja ratkaisemattomia ongelmia, joiden vuoksi hakukysymyksen esittämisellä puhuen ei päästä riittävän hyvään tulokseen. (Colineau & Halber 1999, 1)

Puhemuotoisen tiedonhaun hankaluudet liittyvät paitsi puheentunnistuksen yleisiin ongelmiin (sanojen erilaiset ääntämykset, epäselvä puhe, taustahälyn karsiminen yms.) myös hakukysymyksen jäsennykseen. Kun tiedonhakija ilmaisee hakukysymyksen puhuen, hän ei välttämättä tyydy sanomaan pelkästään hakutermejä, vaan hän voi ilmaista hakukysymyksen esimerkiksi näin:

- Onko sinulla mitään tietoja X:stä?
- Tutkimme X:ää, mitä sinulla on siitä?
- Teen raporttia X:stä, voitko auttaa?
- Haluan videon X:stä.
- Voitko lähettää minulle nauhoja tai printtejä X:stä?

Näin ollen kehittyneen tiedonhakujärjestelmän tulisi pystyä analysoimaan hakukysymys ja ymmärtämään sen perusteella, mitä tiedonhakija oikeastaan hakee (Colineau & Halber 1999, 1-2). Tosin puhekäyttöliittymän kehittämiseen riittänee ensi vaiheessa vähemmänkin kunniahimoinen toteutus. Puhemuotoisessa tiedonhaussa päästään hyvään alkuun, jos käyttöliittymä osaa tulkita edes muutaman keskeisen tiedonhaussa käytetyn käskyn sekä käyttäjän lausuman hakutermin (esim. **etsi *Clinton***, **yhdistä *haut yksi ja kaksi***, **lue *kolmas dokumentti***).

## 6. Puhedokumenttien haun ongelmia

Puhedokumenttien haussa on yhdistettävä kaksi asiaa: tekstimuotoisen tiedon hakutekniikat ja puheen automaattisen tunnistamisen tekniikat. Hakumahdollisuudet riippuvat luonnollisesti siitä, miten dokumentit on indeksoitu. Löytämisen lisäksi tiedonhakujärjestelmän tulee pystyä esittämään dokumentti tiedonhakijalle siinä muodossa, että hän pystyy arvioimaan sen relevanttiuden. Äänimateriaalin tilaavievän luonteen vuoksi tähän ei ole yksinkertaista ratkaisua. Sitä paitsi on tarpeetonta esittää tiedonhakijalle koko tallennetta, jos häntä kiinnostaa vain pieni osa siitä. Ihannetapauksessa tiedonhakija pystyisi sekä tunnistamaan dokumentin että kuuntelemaan siitä potentiaalisesti kiinnostavia kohtia relevanttiuden arvioimiseksi. (Jones et al. 1997, 193)

Puhedokumenttien haussa kohdataan samoja ongelmia kuin tekstihaussakin. Homonymia aiheuttaa hakuvirheitä niin puhutussa kuin kirjoitetussakin tekstissä. Synonymia puolestaan saattaa estää relevantin dokumentin löytymisen, mikäli hakutermi esiintyy indeksissä eri sanana kuin mitä tiedonhakija käyttää (synonyymina). Myös kirjoitusvirheet ja

kielioppivirheet dokumenteissa ja hakukysymyksissä aiheuttavat virheitä aivan samaan tapaan kuin haettaessa tekstimuotoista tietoa. (Jones et al. 1997, 195)

Näiden vanhojen ongelmien lisäksi puhehaussa on myös koko joukko uusia ongelmia. Kahden eri sanan lausuminen samalla tavalla - homofonia - aiheuttaa hankaluuksia, joista tekstihaussa ei tarvitse piitata (esim. *sadettakin/sadetakin, whales/Wales*). Ongelmia aiheuttaa myös sanojen väärin kuuleminen: puheentunnistusohjelma voi tulkita sanan väärin, mikä saattaa aiheuttaa hakuvirheen. Jos avainsana tulkitaan väärin, se jää löytymättä silloin kun pitäisi ja löydetään kun ei pitäisi. Jos taas väärä sana tulkitaan avainsanaksi, saattaa tulosjoukkoon päästä epärelevantti dokumentti. (Jones et al. 1997, 195) Vastaavanlaiset ongelmat ovat tuttuja tekstihausta, mutta puhehaussa niiden ratkaiseminen on vielä vaikeampaa.

Tiedonhaun päätteeksi tiedonhakijan pitäisi pystyä päättelemään löytämiensä dokumenttien relevanttius ilman, että hänen täytyisi kuunnella koko puhedokumenttia läpi. Tähänkään ei ole valmista ratkaisua, ja myös relevanttiuden määrittely puhehaussa on avoin kysymys.

## 7. Puhehaun tutkimuksia

Puhehaku on poikkitieteellinen tutkimusalue, joka vaatii sekä tietojenkäsittelyn, tiedonhaun että kielitieteen tuntemusta.

Puhehaku on suhteellisen uusi alue sekä tiedonhaun tutkimuksessa että puheen automaattisen tunnistuksen tutkimuksessa. Tehdyt tutkimukset käsittelevät lähinnä puhedokumenttien indeksoinnin ja käyttöönasettamisen ongelmia. Puhehaun tutkimuksissa on käytetty usein samoja pohjaratkaisuja kuin tekstihaussa. Kiinnostus aihetta kohtaan on kasvamassa, ja uusia tutkimusaiheita on noussut esiin. Keskeisiä tutkimuskohteita puhehaussa ovat mm. puhedokumentin indeksointi, puheentunnistus, dokumentin ja kyselyn vastaavuuden mittaaminen sekä dokumentin relevanttiuden määrittely ja relevanssipalautteen toteuttaminen. (Wang 1998) Puhehaun tulosten onnistumista on usein mitattu vertaamalla tuloksia vastaavan tekstihaun tuloksiin. Hakutuloksia voidaan tietenkin arvioida myös perinteiseen tapaan laskemalla hakujen saanti ja tarkkuus.

Yhdysvalloissa Standardien ja tekniikan kansallinen instituutti (*the National Institute of Standards and Technology*) on koko 90-luvun ajan rahoittanut TREC-konferenssia (TREC = *Text REtrieval Conference*), jonka yhtenä tutkimuskohteena on puhedokumenttien haku. Tämän projektin tavoitteena on kehittää sisältöperusteisia hakujärjestelmiä puhedokumentteja sisältäviin digitaalisiin arkistoihin. Projektissa hyödynnetään sekä automaattista puheentunnistusta että tiedonhakutekniikoita. (Garofolo et al. 2000, 1-3)

TRECin yhteydessä tehdyt puhedokumenttien hakua koskevat tutkimukset ovat olleet varsin rohkaisevia. Aluksi projektissa tutkittiin vain tunnettujen dokumentin löytymistä pienestä tietokannasta, mutta tutkimuksessa on sittemmin laajennuttu tutkimaan suurienkin puhedokumenttikokoelmien indeksointia ja hakua. Puheentunnistus on sekä tarkentunut että nopeutunut vuosien 1996-1999 aikana, jolloin TRECin puhedokumenttien haku -projekti on ollut käynnissä. Projektin merkittävimpiä tutkimuksia on tehty Cambridgen yliopistossa Englannissa sekä Carnegie Mellonin yliopistossa Pittsburgissa Yhdysvalloissa. (Garofolo et al. 2000)

Cambridgen yliopistossa on ollut käynnissä Video Mail Retrieval -projekti (VMR), joka pyrkii kehittämään sovellusta puhedokumenttien haulle multimediajärjestelmissä. Projektissa on yhdistetty puheentunnistus sekä tekstihakuun perustuvat tiedonhakutekniikat. Vuonna 1997 valmistuneen tutkimuksen aineistona olivat videomuotoiset sähköpostikirjeet. Järjestelmä indeksoi dokumentit 35 ennaltavalitun avainsanan perusteella, jotka se pystyy tunnistamaan. Tutkimuksessa on vertailtu samojen dokumenttien hakua puhe- ja tekstimuodossa. Tutkimustulosten mukaan puhehaun tarkkuudessa päästään jopa 95 %:iin vastaavan tekstihaun tuloksista, jos järjestelmä on ohjelmoitu tunnistamaan puhuja etukäteen. Jos järjestelmä ei tunnista puhujaa, jäävät tulokset 75 %:iin tekstihaun tuloksista. Tutkimus tehtiin rajoitetussa ja suotuisassa käyttöympäristössä, mutta tulosten perusteella puhedokumenttien haku vaikutti kaiken kaikkiaan kehittämisen arvoiselta idealta. (Jones et al. 1997)

VMR-projektissa on tutkittu myös, millainen indeksointimenetelmä tuottaa parhaan tuloksen puhehaussa: yksittäisten avainsanojen poiminta (*word spotting*), laajan sanaston tunnistaminen (*large vocabulary recognition*) vai lattiisipohjainen foneemien skannaus (*phoneme lattice scanning*). Parhaimpiin hakutuloksiin päästiin, kun videokirjeistä tunnistettiin sanoja laajan sanaston pohjalta. (Puheentunnistaja tunnisti 20 000 eri sanaa.) Tällöin kuitenkin osa tiedonhaun kannalta keskeisistäkin termeistä, mm. erisnimet, jäivät tunnistamatta. Kohtalaisen hyviin tuloksiin päästiin myös foneemien skannauksella, mutta parhaimpaan tulokseen päästiin yhdistelemällä eri

menetelmiä. Parhaimmillaan tuloksissa päästiin 85 %:iin vastaavan tekstihaun tuloksista. (Jones et al. 1996)

Puhehaun tutkimusta on jatkettu Cambridgen yliopistossa edelleen. Vuonna 1999 tehdyssä tutkimuksessa aineistona oli 500 tuntia auditiivista uutismateriaalia. Aineisto syötettiin HTK-puheentunnistusohjelmalle, joka käy aineistoa läpi 13-kertaisella nopeudella alkuperäiseen puhenopeuteen verrattuna. Ohjelman sanavirhetaso oli 20,5 %. Kun tällä menetelmällä indeksoituun tietokantaan tehtiin hakuja, 55,3 % löydetyistä dokumenteista oli relevantteja. Tosin jos aineistosta ei ollut etukäteen manuaalisesti karsittu mainoskatkoja ym. taukoja, jäi hakujen keskimääräinen tarkkuus 41,5 %:iin. (Johnson et al. 1999)

Carnegie Mellonin yliopistossa on vuodesta 1994 asti ollut käynnissä Informedia-projekti, jonka yhtenä osana on tutkittu puheentunnistusta uutisvideoista sekä puheentunnistuksen hyödyntämistä videodokumenttien indeksoinnissa digitaaliseen videotietokantaan. Yliopistolla kehitettyyn Informedia-videokirjasto-järjestelmään liittyy myös puhekäyttöliittymä, mutta hakukysymyksiä voi esittää kirjoittamallakin. Vuonna 1997 tehtyjen testien mukaan puheentunnistuksen taso vaihtelee dramaattisesti sen mukaan, minkä tyyppisestä videomateriaalista tunnistus on tehty. Parhaiten puheentunnistus onnistuu, jos se tehdään yksittäisen uutistenlukijan tai juontajan äänestä. Tällöin vain noin 20 % sanoista tulkitaan virheellisesti. Sen sijaan runsaasti taustahälyä sisältävien ohjelmien (esim. mainokset ja dokumenttiohjelmat) sisältämän puheen sanoista peräti 75-85 % tulkitaan väärin. Kuitenkin tämänkin tunnistamistason katsottiin riittävän videodokumentin indeksoinnin ja haun tarpeisiin. (Hauptman & Witbrock 1997)

Informedia-projektinkin tutkimuksia on sittemmin jatkettu. Vuonna 1999 Informedia-videokirjasto-järjestelmä hyödynsi puheentunnistuksessa Carnegie Mellonin yliopistossa kehitettyä Sphinx-III -puheentunnistusohjelmaa, joka tunnistaa 20 000 eri sanaa. Testien mukaan Sphinx-III:n keskimääräinen sanavirhetaso on 24 %, jos aineisto käydään läpi useampaan kertaan. Jos puheentunnistus halutaan suorittaa nopeammin, on sanavirhetaso 34 %. Silloin kun puheaineisto on selkeää ja helposti käsiteltävää, voidaan Sphinx-III:lla päästä jopa alle 20 % sanavirhetason. (Hauptman & Olligschlaeger 1999)

Puhehakua on tutkittu myös muiden kielien kuin englannin osalta. Zürichin teknillisessä instituutissa Sveitsissä on rakennettu puhedokumenttien hakujärjestelmää, josta voi hakea saksankielisiä puhedokumentteja käyttäen joko saksan- tai ranskankielisiä hakukysymyksiä. Tätä puhujasta riippumatonta järjestelmää on testattu käyttäen aineistona 30 tuntia saksankielistä uutismateriaalia. Järjestelmä indeksoi puhedokumentit foneemiperusteisesti. Alustavien tutkimustulosten mukaan tällainen eri kielten sekoit-taminen ei huononna puhehaun tuloksia merkittävästi, vaikka se tuokin uusia haasteita. (Sheridan et al. 1997)

Edellämainitut tutkimukset on siis tehty pääasiassa englanninkielisellä materiaalilla. Puhehaun tutkimus on kuitenkin voimakkaasti kielisidonnaista. Suomessa puhehakua ei ole vielä tutkittu, eikä suomen kielen aiheuttamista mahdollisista erityisistä eduista tai ongelmista ole siten tietoa. Voidaan kuitenkin olettaa, että suomen sanojen taipuminen ja siitä seuraava sanamuotojen runsaus aiheuttaa ongelmia puheentunnistukselle. Suomessa substantiivista voidaan tuottaa teoriassa noin 2000 ja verbistä peräti 12 000 erilaista taivutusmuotoa, kun taas englannissa sekä substantiivi että verbi voivat esiintyä vain neljässä eri taivutusmuodossa (Karlsson & Koskenniemi 1984, 42). Toisaalta suomen intonaation puute (sanojen lausuminen ja kirjoittaminen samalla tavalla) saattaa helpottaa sanojen ja foneemien tunnistamista. Foneemipohjainen tunnistusmenetelmä vaikuttaisi siten järkevimmältä suomenkielisen puheen tunnistamiseen. Mahdollisesti suomenkielisen puhehaun tulee perustua täysin erilaisille pohjaratkaisuille kuin muiden kielien.

Hyväksytty julkaistavaksi 2.10.2000.

## LÄHTEET

Abberley, D., Kirby, D., Renals, S., Robinson, T. 1999: The THISL broadcast news retrieval system. (http://svr-www.eng.cam.ac.uk/~ajr/esca99/Abberley.pdf)

Cole, R., Mariani, J., Uszkoreit, H., Zaenen, A., Zue, V. 1996: Survey of the state of the art in human language technology. (http://cslu.cse.ogi.edu/HLTsurvey/)

Colineau, N., Halber, A. 1999: A hybrid approach to spoken query processing in document retrieval system. (http://svr-www.eng.cam.ac.uk/~ajr/esca99/Colineau.pdf)

Ferrieux, A., Peillon, S. 1999: Phoneme-level indexing for fast and vocabulary-independent voice/voice retrieval. (http://svr-www.eng.cam.ac.uk/~ajr/esca99/Ferrieux.pdf)

Garofolo, J.S., Auzanne, C.G.P., Voorhees, E.M. 2000: The TREC spoken document retrieval track: A success story. In book: RIAO'2000 Conference proceedings: Content-based multimedia information access. Paris: C.I.D. pp. 1-20. ISBN 2-905450-07-X

Glavitsch, U., Schäuble, P. 1992: A system for retrieving speech documents. In book: SIGIR 92. Proceedings of the fifteenth annual international ACM SIGIR conference on research and development in information retrieval. Copenhagen: ACM. pp. 168-176. ISBN 0-89791-523-2

Hauptman, G., Olligschlaeger, A.M. 1999: Using location information from speech recognition og television news broadcasts. (http://svr-www.eng.cam.ac.uk/~ajr/esca99/Hauptmann.pdf)

Hauptman, G., Witbrock, M.J. 1997: Informedia: News-on-Demand multimedia infromation acquisition and retrieval. In book: Maybury, M.T. (ed.): Intelligent multimedia information retrieval. Menlo Park: AAAI Press. The MIT Press. pp. 215-239. ISBN 0-262-63179-2

Johnson, S.E., Jourlin, P., Spärck Jones, K., Woodland, P.C. 1999: Spoken document retrieval for TREC-8 at Cambridge University. (http://svr-www.eng.cam.ac.uk/~sej28/papers/trec99/)

Jones, G.J.F., Foote, J.T., Spärck Jones, K., Young, S.J. 1996: Retrieving spoken documents by combining multiple index sources. In book: SIGIR 96. Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval. Hartung-Gorre verlag konstanz, Germany. pp. 30-39. ISBN 3-89191-999-9

Jones, G., Foote, J., Spärck Jones, K, Young, S. 1997: The video mail retrieval project: Experiences in retrieving spoken documents. In book: Maybury, M.T. (ed.): Intelligent multimedia information retrieval. Menlo Park: AAAI Press. The MIT Press. pp. 191-214. ISBN 0-262-63179-2

Karlsson, F., Koskenniemi, K. 1984. Koneen kieleksi äidinkieli. Tiede 2000, 1. s. 39-46.

Mani, I., House, D., Maybury, M., Green, M. 1997: Towards content-based browsing of broadcast news video. In book: Maybury, M.T. (ed.): Intelligent multimedia information retrieval. Menlo Park: AAAI Press. The MIT Press. pp. 241-258. ISBN 0-262-63179-2

Oard, D. 1998: Bringing Star Trek to Life: Computers that Speak and Listen. (http://www.clis.umd.edu/~rba/COURSES/f98/795/speech/tsld001.htm)

Oard, D. 2000: Speech Retrieval Resources (http://www.clis.umd.edu/dlrg/speech/)

Räihä, K-J. 1999: Kohti HALin vuotta 2001: Näkevät, kuulevat ja tuntevat tietokoneet. Tietojenkäsittelytiede, kesäkuu 1999. s. 18-22.

Schäuble, P. 1997: Multimedia information retrieval: Content-based information retrieval from large text and audio databases. Boston: Kluwer Academic Publishers. 198 p. ISBN 0-7923-9899-8

Sheridan, P., Wechsler, M., Schäuble, P. 1997: Cross-language speech retrieval: Establishing a baseline performance. In book: Proceedings of the 20th annual international ACM SIGIR conference on research and development in information retrieval. Philadelphia: ACM. pp. 99-108. ISBN 0-89791-836-3

Wang, H. 1998: Speech Retrieval (http://wwwckip.iis.sinica.edu.tw/conference/HANDOUT/Hsin-minWang/index.htm)

## Tämän numeron kirjoittajat:

Järvelin, Kalervo, professori, Tampereen yliopisto

Kekäläinen, Jaana, YTT, tutkija, Tampereen yliopisto

Lehtokangas, Raija, YTM, Tampere

Leppänen, Erkka, YTM, Tampere

Saarti, Jarmo, FT, Kuopio