Multimodaalisten syötteiden käsittely: modaliteettien käyttö ja fuusiomenetelmät

Tanja Malmberg

Tampereen yliopisto
Tietojenkäsittelytieteiden laitos
Tietojenkäsittelyoppi
Pro gradu -tutkielma
Ohjaaja: Roope Raisamo
Kesäkuu 2007

Tampereen yliopisto
Tietojenkäsittelytieteiden laitos
Tietojenkäsittelyoppi
Tanja Malmberg: Multimodaalisten syötteiden käsittely: modaliteettien käyttö ja fuusiomenetelmät
Pro gradu -tutkielma, 66 + 3 sivua
Kesäkuu 2007

Multimodaaliset järjestelmät ovat yritys muuttaa ihmisen ja tietokoneen välistä vuorovaikutusta luonnollisempaan suuntaan tarjoamalla uusia tapoja kommunikoida tietokoneiden kanssa. Syöte voidaan antaa useiden eri vuorovaikutuskanavien kautta, eikä käyttäjien tarvitse pitäytyä teennäisissä kommunikaatiotavoissa, joita perinteiset käyttöliittymät suosivat. Toisaalta multimodaalisuus tuo mukanaan useiden syötteiden tulkitsemiseen ja yhdistämiseen liittyvät ongelmat, joiden ratkaiseminen on edellytys tehokkaalle ja luotettavalle multimodaaliselle vuorovaikutukselle.

Tässä tutkielmassa käsitellään modaliteettien käyttöön liittyviä erityispiirteitä ja tutkimustuloksia, jotka tulee ottaa huomioon suunniteltaessa eri modaliteeteilta tulleiden syötteiden yhdistämistä. Ennen varsinaisia fuusiomenetelmiä fuusioprosessia käsitellään yleisluontoisesti: sen tasoja, toteutusta ja ongelmia. Syötteiden käsittely on multimodaalisissa järjestelmissä keskeisessä asemassa, sillä tulkinnan onnistuminen pitkälti määrittää järjestelmän toimintavarmuuden. Fuusiomenetelmien toteuttamistekniikoiden lisäksi niiden yhteydessä pyritäänkin käsittelemään syöteongelmien ratkaisemiseksi kehiteltyjä virheiden käsittelytekniikoita. Tutkielmassa käydään läpi useita erilaisia fuusiotekniikoita, joista jokainen pyrkii ratkaisemaan multimodaalisten syötteiden yhdistämiseen liittyvät hankaluudet omalla tavallaan. Lopuksi esitellään vielä lyhyesti EMMA-merkintäkieli, joka on tiedonsiirtoformaatti multimodaalisia järjestelmiä varten. Lisäksi se tarjoaa useita elementtejä ja attribuutteja ohjaamaan ja avustamaan multimodaalisten syötteiden yhdistämistä.

Avainsanat ja -sanonnat: multimodaalisuus, modaliteetti, fuusiomenetelmät, EMMA.

Sisällys

1. Johdanto

Ihmisen ja tietokoneen välinen vuorovaikutus on pitkään rajoittunut perinteisten käyttöliittymien tarjoamiin vähäisiin keinoihin kommunikoida tietokoneen kanssa. Myös tietokone on osaltaan joutunut esittämään informaation tavalla, joka ei käyttäjän kannalta välttämättä ole paras mahdollinen. Vuorovaikutus on tavanomaisesti käyttäjän osalta ollut lähinnä komentojen antamista hiiren ja näppäimistön avulla sekä palautteen saamista visuaalisesti tietokoneen näytön kautta. Erityisryhmien kohdalla tämä on vaikeuttanut tai jopa estänyt tietokoneiden käytön ja toisaalta osoittautunut epäkäytännölliseksi uusien tietokonelaitteiden, kuten matkapuhelimien, kohdalla. Vuorovaikutusongelmien ratkaisemiseksi on esitetty multimodaalisia käyttöliittymiä, jotka sallivat useiden aistien hyödyntämisen kommunikaation parantamiseksi. Ne ovat yksi askel kohti luonnollisempien käyttöliittymien kehitystä.

Multimodaaliset järjestelmät ovat verrattain uusi suuntaus ihmisen ja tietokoneen vuorovaikutuksessa ja kehitystyö on niiden osalta vielä vilkasta. Ensimmäinen yritys yhdistää useiden aistien kautta tapahtuva viestintä eli eri modaliteeteilla annetut syötteet tehtiin jo vuonna 1980, kun Bolt [1980] esitteli tunnetun *Put-that-there*-järjestelmänsä. Se salli käyttäjien yhdistää puhesyötteen komennot ja osoituseleet hallitakseen seinälle heijastettua maailmaa kuvaavaa projektiota. Käyttäjät pystyivät muun muassa luomaan ja siirtämään objekteja käyttämällä pronomineja kuten ”that” ja ”there”, jotka osoituseleestä riippuen viittasivat joko objekteihin itseensä tai niiden sijaintiin. Sittemmin multimodaalisten käyttöliittymien kehittyminen on ollut nopeaa ja uusia modaliteetteja on otettu mukaan sovelluksiin.

Multimodaalisessa vuorovaikutuksessa modaliteettien käyttö nousee tärkeään asemaan: niiden valinta ja yhdistely vaikuttavat siihen, miten käyttäjän antamat syötteet tulee käsitellä. Usein käyttäjien multimodaaliset vuorovaikutustavat ovat yhteydessä heidän toimintaansa luonnollisessa viestintätilanteessa. Modaliteettien käyttöä käsitteleviä tutkimustuloksia (esimerkiksi Oviatt and Olsen [1994]) voidaankin käyttää ohjaamaan multimodaalisten järjestelmien kehittämistä kohti ihmisten välistä luonnollista kommunikaatiota ja joustavampia toimintamalleja. Lisäksi erilaisten vuorovaikutustapojen tunteminen auttaa helpommin hallitsemaan useiden erilaisten syötteiden yhdistämistä, joka on multimodaalisten järjestelmien ominaispiirre. Syötteiden yhdistämistä on lähestytty useasta näkökulmasta alkaen aina fuusioprosessin määrittelystä [Salber *et al*., 1995] itse fuusiotekniikan yksityiskohtiin [Johnston and Bangalore, 2000].

Vaikka multimodaaliset järjestelmät tarjoavatkin perinteisiin käyttöliittymiin verrattuna tehokkaamman ja luonnollisemman vuorovaikutuksen, on niiden käytännön toteutuksissa vielä useita ongelmia ratkaistavaksi. Sovelluskehittäjille ne tuovat uusia haasteita: multimodaaliset järjestelmät ovat huomattavasti monimutkaisempia kuin perinteiset käyttöliittymät ja näin ollen niiden suunnittelu ja toteutus vaikeampaa. Modaliteetti-

en vaatimat vuorovaikutuslaitteet ovat vasta alkutekijöissään, eikä alan tietotaitokaan ole vielä yleistynyt samalla tavalla, kuin perinteisten järjestelmien kohdalla. Keskeisen haasteen tarjoaa eri modaliteeteilta tulleiden syötteiden synkronisaatio ja integraatio, joka ei ole yksinkertainen ongelma ratkaistavaksi, kun huomioon otetaan käyttäjien kesken yksilöllisesti vaihteleva vuorovaikutustapa ja syötteiden mahdollinen virheherkkyys.

Kuitenkin multimodaalisen järjestelmän on pystyttävä tulkitsemaan käyttäjän antamat syötteet oikein, jotta se voi toimia annettujen komentojen edellyttämällä tavalla. Riittävän toimintavarmuuden saavuttamiseksi tulee huomioida useita näkökohtia: esimerkiksi syötteiden uni- tai multimodaalisuus, modaliteettien keskinäinen yhteistyö sekä syötteisiin itseensä liittyvät monenlaiset ominaispiirteet. Tulkintojen oikeellisuus johtaa onnistuneeseen syötteiden yhdistämiseen, jonka tuloksena tietokone saa yksiselitteisen syötteen käsiteltäväkseen. Toisaalta syötteiden onnistunut yhdistäminen vaatii taakseen myös sovellusalueeseen sopivan ja tehokkaan fuusiomenetelmän, joka voidaan valita olemassa olevien ratkaisujen joukosta tai toteuttaa itse. Tärkeintä on varmistaa fuusion onnistuminen ilman, että syötteiden tulkintojen merkitys muuttuu tai että niiden väärinymmärtämisen riski on liian suuri.

Tämän tutkielman tavoitteena on tarkastella multimodaalisten järjestelmien syötteiden käsittelyä alkaen aina yksittäisten modaliteettien käytöstä päättyen syötteiden yhdistämisestä vastaavien fuusiomenetelmien arvioimiseen. Multimodaalisuus tuskin tulee katoamaan ihmisen ja tietokoneen vuorovaikutuksen tutkimuksesta, vaan on tulevaisuudessa yksi merkittävimmistä ratkaisuista luonnollisia käyttöliittymiä suunniteltaessa. Vaikka tavallinen käyttäjä ei vielä täysin pystykään hyötymään kaikista alan toteutuksista, ovat useat sovellus- ja laitekehittäjät osoittaneet mielenkiintoa multimodaalisuuteen ja tuoneet markkinoille sitä hyödyntäviä toteutuksia (esimerkiksi tuntopalautetta käyttävät hiiret ja puhesyötteiden antamisen mahdollistavat sovellukset). Aihepiiriä käsittelevien aiempien tutkimusten ja toteutusten tarkasteleminen voikin auttaa välttämään tai vaihtoehtoisesti ratkaisemaan niitä ongelmia, joita multimodaalisten järjestelmien kohdalla tulee huomioida.

Seuraavassa luvussa aloitetaan multimodaalisten järjestelmien määrittelyllä, jossa käydään läpi alueeseen liittyvät peruskäsitteet selityksineen. Samassa yhteydessä tarkistellaan lisäksi multimodaalisuuteen liittyviä hyviä ja huonoja puolia, jotka erottavat multimodaaliset järjestelmät unimodaalisista ja perinteisistä käyttöliittymistä. Luvussa 3 käsitellään modaliteettien käyttöä. Ensin käydään läpi sovelluksissa yleisimmin käytettyjä modaliteetteja, jonka jälkeen perehdytään niiden yhdistelytapoihin. Modaliteettien yhteistyöstä on olemassa useita tutkimustuloksia, joita voidaan hyödyntää järjestelmän modaliteetteja valittaessa. Luvussa käsitellään myös syötteiden synkronisaatiota eli tahdistusta: käyttäjät antavat pääasiassa joko samanaikaisia tai jaksoittaisia syötteitä. Oman

lisänsä tuo käyttäjien omaksuma vuorovaikutustapa, josta saatuja tutkimustuloksia käydään läpi viimeiseksi.

Luvussa 4 perehdytään multimodaalisten syötteiden yhdistämiseen yleisellä tasolla. Tarkastelu aloitetaan fuusioprosessin määrittelyllä ja sen suoritustasojen esittämisellä. Toteutuksen osalta mukaan otetaan arkkitehtuurit ja agentit, jotka ovat merkittävässä osassa multimodaalisia järjestelmiä ja niiden fuusioprosessia suunniteltaessa. Lopuksi käsitellään syötteisiin liittyviä erinäisiä ongelmia ja niiden ratkaisumenetelmiä. Luvussa 5 fuusioprosessia tarkastellaan konkreettisemmin: luvussa esitellään useita eri tekniikoihin pohjautuvia fuusiomenetelmiä. Menetelmien yleisen kuvauksen lisäksi huomiota kiinnitetään siihen, miten ne yrittävät ratkaista luvussa 4 kuvattuja syötteiden käsittelyn ongelmia.

Multimodaalisten syötteiden merkintätavoista lähempään tarkasteluun pääsee W3C:n kehittämä merkintäkieli EMMA (Extensible MultiModal Annotation), joka tukee syötteiden yhdistämisen kuvaamista. Luvussa 6 käydään läpi EMMAn osuutta multimodaalisessa viitekehyksessä ja sovelluksissa, jotka jo hyödyntävät EMMAn keskeneräisiä määrittelyjä. Luvussa esitellään pääpiirteissään EMMA-dokumentin rakenne ja siihen kuuluvat elementit ja annotaatiot, joiden tarkempi kuvaus on löydettävissä W3C:n esittämästä teknisestä dokumentista [Johnston *et al*., 2007]. Luku päättyy lyhyeen esimerkkiin, joka havainnollistaa EMMAn käyttöä puhe- ja elesyötteiden kanssa. Luvussa 7 arvioidaan esitettyjen asioiden vaikutusta käytännön sovellussuunnitteluun sekä pohditaan multimodaalisten järjestelmien tulevaisuuden suuntauksia.

2. Multimodaaliset järjestelmät

Multimodaalisuus eli moniaistisuus on yksi uusista suuntauksista tietojenkäsittelytieteessä, joka on herättänyt runsaasti kiinnostusta sekä tutkijoiden että sovelluskehittäjien parissa. Ihmisen ja tietokoneen välisessä vuorovaikutuksessa termillä tarkoitetaan mahdollisuutta käyttää useampaa samanaikaista kanavaa, joilla kommunikointi osapuolten välillä on mahdollista. Kontekstista ja käyttäjästä riippuen näitä kanavia eli modaliteetteja ovat esimerkiksi näkö, kuulo ja tunto. Tavoitteena on multimodaalisuuden avulla parantaa ihmisen ja tietokoneen vuorovaikutusta muuttamalla sitä ihmisten välisen luonnollisen kommunikaation suuntaan.

2.1. Multimodaalisuus

Jo termin ensimmäinen osa ”multi” kertoo, että käytettävissä on useampi kuin yksi modaliteetti samanaikaisesti. Sen sijaan modaliteetin käsite on vaihdellut määrittelijästä ja aihealueessa riippuen. Nigay ja Coutaz [Nigay and Coutaz, 1993] määrittelevät sen kommunikaatiokanavaksi, jonka kautta voidaan välittää tai hankkia informaatiota. Modaliteetti on tiiviisti yhteydessä ihmisaisteihin ja kattaa tavan, jolla idea esitetään tai havaitaan. Toisin sanoen, ihmisen ja tietokoneen vuorovaikutuksessa sillä tarkoitetaan eri aistien kautta tapahtuvaa viestintää. Eniten kiinnostusta ovat herättäneet visuaalinen, auditiivinen, haptinen ja tasapainon modaliteetti.

Ihmisten keskinäinen vuorovaikutus on luonnostaan multimodaalista. Kommunikoidessaan ihmiset käyttävät hyväkseen puheen lisäksi niin sanottua sanatonta viestintää, johon kuuluvat esimerkiksi eleet, jäljittely, äänen variaatiot ja kasvojen liikkeet. Näin he hyödyntävät useita viestintäjärjestelmiä eli modaliteetteja, joiden avulla haluttu viesti välitetään vastaanottajalle useampaa aistia käyttäen [Bunt, 1998]. Eri modaliteetteja ei käytetä itsenäisesti toisistaan riippumattomina, vaan ne usein joko vahvistavat toisiaan viestin oikean merkityksen välittämiseksi tai viestin merkitys on jakautunut useiden modaliteettien kesken, jolloin vastaanottajan on pystyttävä yhdistämään sekä kielellinen että ei-kielellinen osa [De Angeli *et al.*, 1998]. Viestitetyn tiedon merkitys on siis riippuvainen enemmän tai vähemmän samanaikaisesti käytetystä toisesta tai useammasta modaliteetista.

Bunt [1998] huomauttaa ihmisten käyttävän luonnollisessa kommunikoinnissaan niin kutsuttua maksimointiperiaatetta (*multimax principle*). Tällöin viestintätilanteen molemmat osapuolet pyrkivät käyttämään kaikkia niitä modaliteetteja ja kanavia, jotka ovat sillä hetkellä mahdollisia. Maksimaalinen käyttö lisää viestinnän tehokkuutta sekä vähentämällä toistoa että antamalla mahdollisuuden esittää samanaikaisesti useita informaation eri puolia, jolloin viestijä voi maksimaalisesti vaikuttaa toiseen osapuoleen. Lopullinen viestinnän onnistuminen ei kuitenkaan ole kiinni siitä kuinka monta modali-

teettia ja kanavaa on käytössä, vaan siitä pystyvätkö osapuolet käsittelemään kaikkea niiden tarjoamaa tietoa.

Ihmisen ja tietokoneen välistä vuorovaikutusta rajoittaa se, että siinä maksimoinnin periaate ei toteudu: vuorovaikutuksesta puuttuvat useat ihmisten välisessä viestinnässä käytetyt modaliteetit tai ne ovat mukana hyvin yksinkertaisessa muodossa [Bunt, 1998]. Esimerkiksi keskusteleminen tavanomaiseen tapaan ei ole mahdollista, vaikka nykyiset puhesovellukset pystyvätkin hyvin tunnistamaan erilaisia sanontoja ja kasvojenliikkeitä tarkkailemalla on voitu tehdä päätelmiä käyttäjän emotionaalisesta tilasta. Tietokoneiden kanssa toimiessaan käyttäjät ovat pakotettuja omaksumaan rajallisen modaliteettien käytön, joka ei vastaa laisinkaan elävän elämän viestintätilanteita.

Multimodaalisten käyttöliittymien tavoitteena on parantaa ihmisen ja tietokoneen välistä vuorovaikutusta. Uusia vuorovaikutuskanavia lisäämällä tietokoneesta pyritään luomaan yhä ihmismäisempi, jotta kommunikaatio koneen ja ihmisen välillä muuttuisi enemmän luonnollisen viestinnän suuntaan. Oviatt *et al.* [2000] toteavat multimodaalisten käyttöliittymien olevan helpompia oppia ja käyttää. Useat käyttäjät myös suosisivat niitä sovelluksissa yli perinteisten käyttöliittymien. Heidän mukaansa multimodaalisuus mahdollistaisi tietotekniikan laajentumisen yhä vaativammille sovellusalueille, kasvattaisi käyttäjäkuntaa antamalla yhä useammalle mahdollisuuden käyttää tietokonetta ja mukautuisi nykyratkaisuja helpommin hankaliin käyttötilanteisiin. Tulevat käyttöliittymät hyödyntäisivät huomaamattomampia, joustavampia, tehokkaampia ja vahvempia ilmaisukeinoja, joihin ihmisten viestintä tyypillisesti perustuu.

Ihmisen ja tietokoneen vuorovaikutus on pitkään ollut riippuvainen perinteisten käyttöliittymien rajoituksista, jotka antavat käyttäjälle vain muutamia peruskeinoja, joilla kommunikoida koneen kanssa. Tämä on hankaloittanut tai jopa estänyt joidenkin erityisryhmien tietokoneen käytön. Toisaalta markkinoille on myös tullut uusia tietokonesovelluksia sisältäviä laitteita, joita usein käytetään liikkuessa, jolloin käsien käyttö saattaa olla mahdotonta, ja joissa laitteen pienen koon vuoksi perinteiset tiedonsyöttömahdollisuudet ovat rajoittuneet. Koska ihmiset suosivat luonnollisten viestintäkeinojen käyttöä informaation vaihdossa, voivat useat modaliteetit yhdessä tarjota vuorovaikutustavan, joka kirkkaasti peittoaa perinteisen hiiri ja näppäimistö -vuorovaikutusmallin.

Multimodaaliset käyttöliittymät mahdollistavat vuorovaikutuksen toteuttamisen perinteisten modaliteettien sijaan jonkin muun, paremmin käyttötilanteeseen soveltuvan modaliteetin avulla. Paternò [2004] esimerkiksi toteaa auditiivisen kanavan olevan parempi yksinkertaisten tai lyhyiden viestien lähettämiseen, tapahtumista ilmoittamiseen, välittömiin toimiin ja käyttäjän liikkuessa, kun taas visuaalinen kanava on hyödyllinen monimutkaisten ja pitkien viestien lähettämisessä, etäisyyksien määrittämisessä, useita toimintoja suoritettaessa sekä meluisissa ympäristöissä. Näin voidaan lisätä sekä viestintänopeutta että tarkkuutta, kun käyttäjälle annetaan mahdollisuus valita sen hetkiseen tilanteeseensa parhaiten sopiva modaliteetti.

Koska puhuttu kieli on tyypillisesti merkittävin osa ihmisten välistä viestintää, on puhe usein keskeisessä roolissa, kun informaatiota välitetään multimodaalisissa käyttöliittymissä ihmisen ja tietokoneen kesken. Myös esimerkiksi kasvojen ilmeitä, silmänliikkeitä ja osoituseleitä on käytetty hyväksi käyttäjän toimintaa tulkittaessa. Oviatt *et al.* [1998] muistuttavatkin, että multimodaalista vuorovaikutusta hyödyntämällä on helppo korvata liian monimutkaiset, yksityiskohtaiset tai ristiriitaiset kielelliset ilmaisut. Tietokone voi vuorostaan simuloida sanatonta viestintää niin kutsutun keinopään avulla, joka on ihmisen päätä ja kasvoja jäljittelevä sovellus.

2.2. Hyötynäkökulmia

Multimodaaliset käyttöliittymät ovat yritys parantaa ihmisen ja tietokoneen välistä vuorovaikutusta. Oviatt [1999a] toteaa niiden antavan käyttäjille enemmän ilmaisuvoimaa, luonnollisuutta, joustavuutta ja liikkuvuutta. Hyvin suunnitellut ja toteutetut multimodaaliset järjestelmät yhdistävät toisiaan täydentävät modaliteetit, jolloin jokaisen vahvuudet pääsevät paremmin esille ja yksittäinen modaliteetti voi kattaa toisen heikkoudet. Näin saavutetaan suurempi toimintavarmuus ja luotettavuus kuin unimodaalisissa järjestelmissä. Multimodaalisuuden hyötyjä ovat muun muassa:

- Yksittäisen modaliteetin aiheuttamat virheet ja ristiriitaisuudet voidaan usein korjata käyttämällä toista modaliteettia. Esimerkiksi modaliteetit voivat keskinäisen kompensaation kautta parantaa syötteen tunnistuksen tarkkuutta. Oviatt [1999a] huomauttaa käyttäjien myös välttävän multimodaalisissa käyttöliittymissä sellaisia syötetapoja, joiden he uskovat olevan virheherkkiä. Samalla he käyttävät yksinkertaisempaa kieltä ja virhetilanteissa vaihtelevat modaliteetteja siten, että voivat ratkaista ongelman mahdollisimman tehokkaasti.
- Oviatt [1999a] toteaa käyttäjillä olevan vahva ja lähes universaali taipumus suosia multimodaalista vuorovaikutusta, joka on heille tuttua ihmisten välisestä viestinnästä. Multimodaalisuus mahdollistaakin jo opittujen kommunikaatiotaitojen hyödyntämisen ihmisen ja tietokoneen vuorovaikutuksessa. Lisäksi käyttäjät voivat tukeutua omaksumaansa vuorovaikutustapaan, jonka on todettu olevan melko pysyvä ja yksilöllisesti vaihteleva.
- Käyttäjät voivat itse valita sopivan modaliteetin tilanteensa mukaan ja vaihdella sitä joustavasti. Näin he voivat välttää yksittäisen modaliteetin liikakäytön [Oviatt, 1999a]. Tämä mahdollistaa tietokoneen käytön myös käyttäjillä, joilla jokin modaliteetti ei ole käytettävissä (esimerkiksi sokeat ja kuurot) ja käyttötilanteissa, joissa perinteiset keinot eivät ole riittäviä tai sopivia (multimodaalisuutta hyödyntämällä voidaan parantaa myös sovellusten tietoturvaa). Samalla multimodaaliset käyttöliittymät antavat käyttäjille enemmän ilmaisumahdollisuuksia: esimerkiksi käytettäessä sekä puhe- että kynäsyötettä voidaan helposti tuottaa erilaisia kuvauksia objek-

teista, tapahtumista, sijainnista tilassa ja näiden keskinäisistä suhteista [Oviatt *et al*., 2000].

- Mahdollisuus valita käytettävä modaliteetti tehostaa työskentelyä. Käyttötilanteeseen nähden oikea modaliteetti nopeuttaa tehtävän suorittamista, vähentää virheiden mahdollisuutta ja lisää työskentelyn tarkkuutta [Maybury, 2002]. Yhdistämällä eri modaliteettien parhaat ominaisuudet voidaan toteuttaa parempi käyttöliittymä kuin mihin yksittäinen modaliteetti pystyisi: esimerkiksi objektien valinta tai osoittaminen on hankalaa puhekäyttöliittymän avulla, kun taas kynällä se onnistuu luontevasti.
- Multimodaaliset järjestelmät ovat usein käyttäjille perinteisiä käyttöliittymiä intuitiivisempia, jolloin niiden käyttökin on helpompi oppia. Ne ovat myös huomaamattomampia, sillä käyttäjän ei enää tarvitse opetella teennäisiä kommunikaatiotapoja. Tämä tekee vuorovaikutuksesta miellyttävämpää. Lisäksi miellyttävyyttä voidaan lisätä esimerkiksi käyttäjän tunnetiloja havainnoimalla: käyttäjän turhautuessa voidaan muuttaa kommunikointitapaa ja lisätä tukea, jotta vuorovaikutus ei kokonaan katkeaisi [Adelhardt *et al*., 2003].
- Tietokoneiden käyttö on pitkään rajoittunut lähinnä paikallaan oleviin pöytätietokoneisiin, joissa tarvetta uusille vuorovaikutuskanaville ei ole ollut. Nykyisin tietokonesovelluksia on monissa muissakin laitteissa, kuten matkapuhelimissa ja musiikkisoittimissa, joilla käyttöympäristö poikkeaa merkittävästi perinteisestä tietokoneen käytöstä. Oviatt *et al*. [2000] huomauttavat, että käyttäjän liikkuvuus (mobiliteetti) voi johtaa niin kutsuttuun hetkelliseen kykenemättömyyteen (*temporary disability*), jolloin yksittäistä modaliteettia ei voida tietyllä hetkellä käyttää. Multimodaaliset järjestelmät mahdollistavat korvaavan modaliteetin käytön, jolloin käyttäjä voi kaikesta huolimatta suorittaa tehtävänsä.

Kaiken kaikkiaan multimodaalisuus oikein toteutettuna vaikuttaa positiivisesti käyttöliittymän käytettävyyteen lisäämällä käytön miellyttävyyttä ja luonnollisuutta, tarjoamalla vaihtoehtoisia vuorovaikutuskanavia ja helpottamalla sekä tehostamalla työskentelyä. Perinteisiin käyttöliittymiin verrattuna ne monipuolistavat ihmisen ja tietokoneen välistä vuorovaikutusta ja lähestyvät saavutettavuuden periaatetta, jonka mukaan jokaisella tulisi olla tasavertainen mahdollisuus päästä käsiksi digitaaliseen informaatioon.

2.3. Toteutuksen ongelmia

Vaikka multimodaalisuus tuokin mukanaan monia parannuksia perinteisiin käyttöliittymiin verrattuna, on käytännön toteutuksissa vielä runsaasti ongelmia. Aivan helppoa ei näiden ongelmien ratkaiseminen ole, sillä matka täysin luonnolliseen ja ihmismäiseen vuorovaikutukseen ihmisen ja tietokoneen välillä on vielä pitkä. Toisaalta voidaan myös kysyä, kuinka pitkälle tietokoneiden inhimillistämisessä kannattaa mennä, koska loppu-

jen lopuksi tietokone on kuitenkin *vain* laite, jonka tarkoituksena on toimia apuvälineenä erilaisia tehtäviä suoritettaessa.

Tekniikka itsessään on usein pullonkaulana uusien järjestelmien ja laitteiden kehittämiselle. Multimodaaliset järjestelmät ovat perinteisiin käyttöliittymiin verrattuna huomattavasti monimutkaisempia ja näin ollen myös niiden suunnittelu ja toteutus on hankalampaa [Oviatt *et al.*, 2000]. Tarvittava tekniikka on vielä pitkälti kehityksen alla ja lisäksi ongelmana on uusien syöte- ja tulostelaitteiden kallis hinta. Tämä hidastaa multimodaalisten käyttöliittymien siirtymistä tutkimuslaboratorioista tavallisen käyttäjän saataville. Kanninen [2003] tiivistää ongelman ytimen yhteen lauseeseen: ”Mitä enemmän modaliteetteja, sitä suurempi riski.”.

Toinen merkittävä multimodaalisten järjestelmien ongelma liittyy syötteiden integraatioon ja synkronointiin: eri syötelaitteiden kautta tuleva informaatio on yhdistettävä yhden merkityksen omaavaksi syötteeksi, jotta tietokone pystyy toimimaan käyttäjän tavoitteen mukaisesti. Modaliteettien kohdalla vaihtelua on siinä miten samansuuntaista informaatiota ne välittävät: joidenkin modaliteettien syöte on helpommin keskenään verrattavissa ja yhdistettävissä kuin toisilta saatu [Oviatt, 1999a]. Tulkintaa vaikeuttavat osaltaan myös syötteet, jotka saattavat olla keskenään ristiriitaisia, kaksiselitteisiä tai puutteellisia. Jo syötteiden tunnistaminen saattaa olla hankalaa: esimerkiksi puheentunnistus perustuu todennäköisyyksiin, jolloin virheet syötteissä ovat mahdollisia ja usein hankalasti korjattavissa [Oviatt, 2000].

Käyttäjien omaksuma vuorovaikutustapa vaihtelee yksilöllisesti multimodaalisia järjestelmiä käytettäessä: käyttäjät esimerkiksi antoivat komennot joko samanaikaisesti tai jaksoittaisesti yhdistäessään puhe- ja kynäsyötteitä keskenään [Oviatt, 1999a]. Yksilölliset vuorovaikutustavat tulisikin ottaa huomioon järjestelmiä suunniteltaessa, jotta syötteiden tunnistaminen uni- tai multimodaalisiksi paranisi. Ongelmallista tämä on, kun otetaan huomioon, että luonnollinen kommunikaatio sisältää usein käytösmalleja ja automaattisia prosesseja, jotka eivät ole tietoisen hallinnan alla [De Angeli *et al.*, 1998].

3. Modaliteettien käyttö

Multimodaalisissa järjestelmissä syöte voidaan antaa useiden eri modaliteettien avulla edellyttäen, että järjestelmä tukee niitä. Kanninen [2003] painottaa modaliteettien valinnalla olevan merkitystä etenkin sen kannalta, millaista informaatiota ja missä muodossa käyttäjä voi välittää ja vastaanottaa. Huomioon tulee ottaa tehtävässä tarvittavan vuorovaikutuksen ja informaation esitystavan vaatimukset, käyttäjän ominaisuudet ja mieltymykset sekä käyttötilanne. Esimerkiksi puhe ja eleet voivat yhdessä luoda käyttöliittymän, joka on tehokkaampi kuin kummankaan modaliteetin käyttö erikseen.

3.1. Modaliteetit

Bernsen [2002] määrittelee modaliteetin informaation vaihtotavaksi ihmisten tai ihmisten ja koneiden välillä, jossa tieto välittyy tietyn median kautta. Median hän määrittelee informaation fyysiseksi ilmenemismuodoksi, jonka ihminen ja kone pystyvät havaitsemaan. Median käsite liittyy läheisesti psykologian määritelmiin aistimodaliteeteista, joita ovat muun muassa näkö ja kuulo. Modaliteetti on se vuorovaikutuskanava, jolla fyysisesti eri tavoin ilmennetty informaatio välitetään. Tämä informaatio voi tulla saman median kautta hyvinkin erilaisissa muodoissa: esimerkiksi merkkiääni ja synteettinen puhe välittyvät saman median kautta, jolloin Bernsen käyttää esityksellisen modaliteetin (*representational modality*) käsitettä erottelemaan ne toisistaan.

Informaation kulkusuunnan perusteella Bernsen [2002] jakaa modaliteetit vastaanotto- (*input*) tai tuottomodaliteeteiksi (*output*). Modaliteetit voivat myös olla molempia samanaikaisesti. Ihmisen näkökulmasta tarkasteltuna vastaanottomodaliteetit ovat niitä, joiden kautta ihminen vastaanottaa informaatiota ja tuottomodaliteetit taasen niitä, joiden kautta ihminen tuottaa informaatiota. Eri modaliteettien tuottama informaatio välitetään tietokoneelle erilaisten syötelaitteiden kautta. Koska tämän työn painotus on multimodaalisissa syötteissä, jätetään tietokoneen antamat tulosteet tässä käsittelemättä.

Larson [2005] jakaa käyttäjän syötteet kahteen ryhmään: koodattuihin ja tunnistettaviin syötteisiin. Koodattu syöte käsittää esimerkiksi näppäimistön, hiiren ja peliohjaimen kautta annetun syötteen: hiiren painikkeen tuottama tapahtuma tai liike näytöllä koodataan merkkijonoiksi tai näyttökoordinaateiksi. Tunnistettava syöte sisältää muun muassa puheen, käsinkirjoituksen ja katseen. Näiden käyttö vaatii niihin erikoistuneet alajärjestelmät, jotka sitten muuttavat syötteen merkkijonoiksi. Tunnistettavat syötteet ovat huomattavasti virheherkempiä kuin koodatut: parhaimmatkin puheentunnistusjärjestelmät tekevät tunnistusvirheitä 3-5 %:ssa annetuista syötteistä. Modaliteetit, syötetavat ja tietokoneen vastaanottolaitteet esitellään taulukossa 1.

Aisti	Modaliteetti	Syöte	Vuorovaikutuslaite
Näkö	Visuaalinen	Katseenliikkeet	Kamera
Kuulo	Auditiivinen	Puhe, äänet	Mikrofoni
Tunto	Haptinen	Kosketukset, liike, eleet	Paikkasensorit, paineherkät alustat
Haju	Olfaktorinen	Ominaistuoksu	Keinonenä
Maku	Gustatoorinen	-	Keinokieli, keinonenä
Tasapaino	Vestibulaarinen	-	Tasapainoanturit

Taulukko 1. Modaliteetit, käyttäjän syötteet ja tietokoneen vuorovaikutuslaitteet.

3.1.1. Visuaalinen

Visuaalinen eli näköaistiin perustuva modaliteetti on yksi useimmiten käytetyistä modaliteeteista, joka on tuttu jo perinteisistä käyttöliittymistä. Graafinen käyttöliittymä sisältää erilaisia elementtejä, kuten ikkunoita, valikoita, tekstiä ja kuvia, joilla vuorovaikutusta ihmisen ja tietokoneen välillä pyritään ohjaamaan. Taidokkaasti suunniteltu käyttöliittymä johdattaa käyttäjän katseen välittömästi merkittäviin elementteihin hyödyntäen visuaalisia tehokeinoja, joita ovat muun muassa elementtien sijoittelu, värien ja kontrastien käyttö sekä erilaiset korotustekniikat.

Käyttäjän silmänliikettä voidaan seurata erilaisten katseenseuratalaitteiden avulla. Laitteet tallentavat silmänliikkeet ja niiden avulla voidaan saada tarkkaa tietoa sekä katseen kulloisestakin sijainnista että liikkeistä näytöllä. Visuaalinen modaliteetti tukee tyypillisesti muita modaliteetteja, mutta puhe- tai liikuntakyvyn vaikeuksista kärsivälle silmien käyttö saattaa olla ainoa kommunikaatiomenetelmä ulkomaailmaan.

3.1.2. Auditiivinen

Auditiivista eli kuuloaistiin perustuvaa modaliteettia pidetään usein virheellisesti tärkeimpänä syötetapana ja se nähdään niin itseriittoisena, että muut modaliteetit ovat vain tarpeettomia lisäkkeitä. Oviatt [1999a] huomauttaa, että yleisesti ottaen tämä ei kuitenkaan ole totta. Multimodaalinen kieli eroaa luonnollisesta kielestä siinä, että lauseet ovat usein lyhyempiä ja rakenteellisesti yksinkertaisempia, jolloin ne eivät pysty samalla tavalla välittämään informaatiota kuin ihmisten välisessä viestinnässä. Puhe ei siis ole muut poissulkeva, ylivoimaista ajallista etua omaava modaliteetti, vaan toiset modaliteetit voivat välittää informaatiota, johon auditiivinen modaliteetti ei pysty.

Puhekäyttöliittymät mahdollistavat puheen ja äänien käytön syötteenä. Laitetasolla voidaan käyttää vastaanottimena perinteistä mikrofonia. Suurin ongelma puheentunnistuksessa on sen epävarmuus, sillä sanojen tunnistus perustuu tyypillisesti todennäköisyyksiin eli niin kutsuttuihin n-best -listoihin. Lisäksi virheiden korjaaminen saattaa olla

hankalampaa kuin perinteisissä käyttöliittymissä, jolloin puheen käyttäminen voi tuntua käyttäjästä turhauttavalta, vaikka kyseessä onkin hyvin luonnollinen viestintäkeino.

3.1.3. Haptinen

Haptinen eli tuntoaistiin perustuva modaliteetti voidaan jakaa taktiiliseen ja kinesteettiseen modaliteettiin. Näistä taktiilinen liittyy kosketukseen ja kinesteettinen enemmänkin liikkeeseen, vartalon ja raajojen asentoihin. Huolimatta siitä, että kosketusta pidetään yhtenä perusaistimuksista, on haptinen modaliteetti jäänyt visuaalisen ja auditiivisen modaliteetin varjoon. Vasta viime vuosina sen parissa on tehty intensiivistä tutkimusta ja tekniikkaa sekä laitteita kehitelty.

Kosketusta voidaan havainnoida erilaisten sensorien avulla ja vastavoimaa käyttämällä simuloida sitä virtuaaliympäristöissä. McLean [2000] huomauttaa haptisen modaliteetin toimivan tehokkaimmin silloin, kun sitä käytetään yhdessä muiden modaliteettien kanssa. Vaikka kosketuksen kautta onkin mahdollista välittää tietynlaista informaatiota tarkemmin kuin muilla modaliteeteilla, on kosketus luonteeltaan summittainen verrattaessa muiden modaliteettien kykyyn välittää informaatiota ja havaita absoluuttisia sekä relatiivisia suhteita.

3.1.4. Muut

Muut taulukossa 1 mainitut modaliteetit – olfaktorinen, gustatoorinen ja vestibulaarinen – ovat pitkälti vielä tutkimusasteella ja käytettävissä vain erikoissovelluksissa. Tulevaisuudessa selviää, tullaanko niitä hyödyntämään multimodaalisissa käyttöliittymissä. Jos näin käy, on mielenkiintoista nähdä, miten ja millaisten laitteiden kautta näiden modaliteettien käyttö tapahtuu.

3.2. Modaliteettien yhdistely

Ihmisten välisessä viestinnässä käytetään samanaikaisesti useita modaliteetteja, joita yhdistellään toisiinsa tarpeen mukaan. Modaliteettien valintaan ja niiden luonnolliseen yhdistämiseen eri tilanteissa vaikuttaa vahvasti suoritettava tehtävä. McKenzie Mills ja Alty [McKenzie Mills and Alty, 1998] toteavat ihmisten ”lähettävän” eri modaliteettien kautta jatkuvasti sekä täydentävää että vahvistavaa informaatiota, jonka avulla aikomuksia voidaan paremmin kommunikoida toiselle osapuolelle. Tällainen informaation tukeminen on tärkeää erityisesti silloin, jos yksittäisellä modaliteetilla välitetty viesti on epäselvä tai vääristynyt.

Samalla McKenzie Mills ja Alty [McKenzie Mills and Alty, 1998] kuitenkin huomauttavat, että nykyisissä tietokoneissa syötetekniikoita ei juuri koskaan käytetä rinnakkain ja erittäin harvoin ne ovat toisiaan vahvistavia. Multimodaaliset järjestelmät sen sijaan tarjoavat useita vaihtoehtoisia vuorovaikutuskanavia, joiden kautta syöte voidaan antaa tietokoneelle. Järjestelmän tukemien modaliteettien valitseminen ei kuitenkaan ole yksinkertainen tehtävä, vaan huomioon pitää ottaa tehtävässä tarvittavan vuo-

rovaikutuksen ja informaation esitystavan asettamat vaatimukset, käyttäjän ominaisuudet ja mieltymykset sekä käyttötilanne [Kanninen, 2003]. Lisäksi on ongelmallista yrittää yhdistää modaliteetteja, jotka eivät koskaan esiinny yhdessä luonnollisessakaan viestintätilanteessa (esimerkiksi kirjoittaminen ja osoituseleet) niiden omien liikerajoitteiden takia [De Angeli *et al.*, 1998].

3.2.1. Yhdistelytavat

Käyttäjien välillä on havaittu olevan suuria yksilöllisiä vaihteluita siinä miten he yhdistelevät eri modaliteetteja. Valittu tapa on usein pysyvä ja nähtävillä jo vuorovaikutuksen alussa, mikä tekee siitä merkityksellisen eri modaliteeteilta tulevien syötteiden onnistuneen integraation kannalta. Usein modaliteettien yhteistyö voidaan luokitella kuuteen tapaan [Martin *et al*., 1998; yhteistyömuotojen suomennokset Kanninen, 2003]:

- Tasa-arvoinen (*equivalence*), jossa sama informaatio voidaan välittää käyttäen useampaa eri modaliteettia. Käyttäjä voi valita tilanteeseensa ja mieltymyksiinsä nähden parhaan modaliteetin vaihtoehtoisten vuorovaikutuskanavien joukosta.
- Vahvistava (*redundancy*), jossa sama informaatio voidaan välittää samanaikaisesti useiden modaliteettien kautta. Tällöin eri modaliteettien kautta tulleet, mutta samaa tavoitettavat tukevat syötteet vahvistavat toisiaan. Jos yhden modaliteetin syöte on riittävä käyttäjän tavoitteen tulkitsemiseen, ei muiden modaliteettien välittämää informaatiota välttämättä tarvita. Vahvistava yhdistämistapa on hyödyllinen varsinkin silloin, jos yksittäisen modaliteetin kautta tullut informaatio muuttuu yllättäen käyttökelvottomaksi, jolloin muiden modaliteettien välittämän, aiemmin käyttämättä jääneen tiedon avulla voidaan onnistuneesti tulkita syöte [McKenzie Mills and Alty, 1998]. Lisäksi vahvistamisesta on hyötyä tehtävissä, joissa virhetulkintojen määrän tulee pysyä pienenä. Multimodaalisesta vuorovaikutuksesta puhuttaessa uskotaan eri modaliteettien välittämän informaation olevan aina toisiaan vahvistavaa. Oviatt [1999a] kumoaa väitteen ja huomauttaa ihmisten keskuudessa vallitsevan luonnollisen vuorovaikutuksen olevan itse asiassa täydentävää, ei vahvistavaa.
- Täydentävä (*complementary*), jossa eri modaliteetit välittävät samaan tavoitteeseen liittyvää, eriävää informaatiota, joka lopussa yhdistetään yhdeksi kokonaisuudeksi. Tämä mahdollistaa sopivimman modaliteetin valinnan tietynlaisen informaation välittämiseen, jota muiden modaliteettien kautta saatu tieto voi täydentää. Oviatt *et al*. [1997] huomasivat tutkimuksessaan, että puhe- ja kynäsyötettä käytettiin johdonmukaisesti tuottamaan erilaista ja toisiaan täydentävää informaatiota. Vuorovaikutustilanteessa tekijää, toimintaa ja kohdetta kuvaavat tiedot lähes aina ilmaistiin puheella, kun taas sijaintia kuvaava tieto kirjoitettiin. Oli myös erittäin harvinaista, että kaikki nämä tiedot toistettiin molemmilla modaliteeteilla. Oviatt [1999a] täsmentääkin, että multimodaalisten järjestelmien suunnittelijoiden ei tulisi liikaa luottaa vahvistavaan informaatioon käsitellessään multimodaalisia syötteitä.

- Erikoistunut (*specialization*), jossa tietynlainen informaatio välitetään aina samalla modaliteetilla. Erikoistunut yhdistämistapa toimii siis vastakohtana tasa-arvoiselle yhdistämiselle. Käyttäjä voi suosia tiettyä modaliteettia tehtävää suorittaessaan muiden modaliteettien sijasta, vaikka tarjolla olisikin enemmän kuin yksi tasa-arvoinen modaliteetti. Erikoistumista voidaan tarkastella myös multimodaalisen järjestelmän kannalta: se antaa suunnittelijalle mahdollisuuden määrittää sopivin modaliteetti kuhunkin tehtävään.
- Modaliteettia muuttava (*transfer*), jolloin yhden modaliteetin välittämä informaatio on toisen modaliteetin käytettävissä ja toimii näin toiminnan käynnistäjänä. Käyttäjän antamat komennot kulkevat eteenpäin eri modaliteettien kautta, kunnes tehtävän tavoite on saavutettu.
- Yhtäaikainen (*concurrency*), jossa eri modaliteetteja käytetään välittämään samanaikaisesti informaatiota. Tämä informaatio on luonteeltaan itsenäistä eikä sitä voi yhdistää muiden modaliteettien välittämän tiedon kanssa. Yhtäaikainen yhdistämistapa saattaa olla käyttäjälle kognitiivisesti rasittava, sillä tarkkaavaisuus joudutaan jakamaan usean modaliteetin kesken.

Modaliteettien yhteistyö määrittää sen, miten multimodaaliset syötteet tulisi tulkita: voidaanko osa niistä jättää huomiotta vai pitääkö kaikki syötteet käsitellä oikean tulkinnan saavuttamiseksi. Vaikka käyttäjän omaksuma vuorovaikutustapa on suhteellisen pysyvä, saattaa modaliteettien yhdistelytapojen vaihtelu auttaa käyttäjää esimerkiksi paremmin sopeutumaan uuteen käyttötilanteeseen, parantamaan syötteiden tulkinnan oikeellisuutta hankalissa olosuhteissa tai siirtämään informaatiota yhdeltä modaliteetilta toiselle. Ongelmia voikin tuottaa se, jos käyttäjä sinnikkäästi pitäytyy tietyssä yhdistelytavassa siitä huolimatta, että se on tilanteen tai tehtävän kannalta tehoton.

3.2.2. Tutkimustuloksia

Seuraavaksi luodaan lyhyt katsaus siihen millaisia tuloksia eri tutkimuksissa on saatu modaliteettien yhdistämisestä ja sen merkityksestä vuorovaikutuksen laadun kannalta. Ensimmäiseksi tarkastellaan puhetta ja kirjoitettua tekstiä eli kynäsyötettä, sitten puhetta ja katsetta sekä lopuksi puhetta ja eleitä. Tarkastelu rajoittuu tässä käyttäjän näkökulmaan eli siihen, miten ihminen käyttää eri modaliteetteja ollessaan vuorovaikutuksessa tietokoneen kanssa.

Oviatt ja Olsen [Oviatt and Olsen, 1994] tutkivat, kuinka ihmiset yhdistävät puhetta ja kirjoitusta multimodaalisen vuorovaikutuksen aikana. Vuorovaikutusta pyrittiin simuloimaan palvelujen välitysjärjestelmän avulla eikä käyttäjän tapaan ilmaista itseään vaikutettu. Oviatt ja Olsen huomasivat käyttäjien suosivan informaation välityksessä puhetta, vaikka kirjoitusta käytettiinkin tietyissä kohdissa vuorovaikutusta: esimerkiksi numerot ja todelliset nimet kirjoitettiin useammin kuin muu teksti. Kaikista puheen ja kirjoituksen yhdistämistavoista 57 %:ssa vaikuttavana tekijänä pidettiin käyttäjien hen-

kilökohtaista mieltymystä valita tietty yhdistäminen tehtävän sisällöstä ja esitystavasta riippuen. Yhtäaikainen modaliteettien käyttö havaittiin harvinaiseksi: kaikista sanoista alle 1 % kommunikoitiin käyttäen sekä puhetta että kirjoitusta.

Luonnollisessa viestinnässä silmät sekä välittävät tunnetiloja että ilmaisevat tarkkaavaisuuden suunnan. Zhang *et al*. [2004] yhdistivät tutkimuksessaan katseen ja puheen käytön. He huomasivat, että jopa lyhyet, yksinkertaiset ja merkitykseltään ristiriitaiset lauseet voivat saavuttaa saman tehon kuin pitkät ja täydelliset lauseet silloin, kun täydentävää informaatiota saadaan katsesyötteen kautta. Täydentävä yhdistämistapa mahdollisti puheentunnistuksen virheiden korjaamisen silmänliikkeiden havainnointia hyödyntämällä ja toisinpäin. Lyhyitä lauseita suosiva multimodaalinen järjestelmä katseenseurannalla varustettuna onkin usein hyvin tehokas.

Kaur *et al*. [2003] huomasivat suuria yksilöllisiä eroja koehenkilöiden välillä tutkiessaan katseen ja puheen yhdistämistä. Tutkimuksessa havaittiin, että katseen siirtäminen tehtävän vaatimaan kohteeseen aloitetaan hyvin suurella todennäköisyydellä ennen puhuttua käskyä, jolloin katse vahvistaa käyttäjän auditiivista modaliteettia. Adelhardt *et al*. [2003] pyrkivät tunnistamaan käyttäjän tunnetiloja yhdistämällä keskenään kasvojen ilmeet, puheen ja eleet. Kaikki ne ovat soveltuvia tunnetilojen tunnistamiseen, joskin vain hyvin harvat ihmiset käyttävät aina jokaista näistä modaliteeteista tunnetilansa näyttämiseen. Adelhardt *et al*. painottavat multimodaalisuudesta olevan erityistä hyötyä silloin, kun jokin luetelluista modaliteeteista puuttuu: tällöin jäljellä olevien avulla voidaan yhä tehdä päätelmiä käyttäjän tunnetilasta.

Yksi suosituimmista kommunikaatiotavoista multimodaalisissa järjestelmissä on puheen ja osoituseleiden yhdistäminen keskenään. Oviatt [1999a] muistuttaa kuitenkin yksinkertaisten, kohteen valintaan tähtäävien eleiden olevan ilmaisuvoimaltaan heikkoja ja vastaavan perinteistä hiiri-metaforaa. Rajoittuminen pelkästään puhu-ja-osoita-vuorovaikutukseen ei anna käyttäjälle kaikkea sitä toiminnallisuutta, jota multimodaalisilla järjestelmillä yritetään saavuttaa. De Angeli *et al.* [1998] tutkivat osoituseleiden ja puheen yhdistämistä tavoitteena parantaa multimodaalisten järjestelmien käytettävyyttä. Koetilanteessa annetuissa käskyissä 11 %:ssa näitä modaliteetteja käytettiin toisiaan vahvistaen, jonka arveltiin johtuvan ihmisten tavasta olla liiankin täsmällisiä tietokoneen kanssa toimiessaan. Tutkimuksessa huomattiin myös osoituseleiden vahvistavan puhetta pääasiassa silloin, kun kielellinen ilmaus oli lyhyt ja tarkka.

McKenzie Mills ja Alty [McKenzie Mills and Alty, 1998] tutkivat modaliteettien toisiaan vahvistavaa yhdistämistä syötteiden virheiden ja ristiriitaisuuksien selvittämiseksi. Tavoitteena oli myös osoittaa, että vahvistavaa yhdistämistapaa voidaan hyödyntää yksinkertaisenkin sovelluksen yhteydessä. Tutkimuksessa puheen ja eleiden yhdistelmän huomattiin parantavan eleiden tunnistamisvarmuutta ja vähentävän näin syötevirheitä. Irawati *et al*. [2006] käyttivät puhetta ja eleitä tutkiessaan multimodaalista vuorovaikutusta lisätyn todellisuuden (*augmented reality*) yhteydessä. Yhdessä nämä moda-

liteetit voivat tehdä käyttöliittymästä hyvin intuitiivisen: puhesyötteen vahvuudet kattavat eleiden rajoitukset ja päinvastoin (ellei vuorovaikutuksessa sitten käytetä pelkkiä osoituseleitä). Irawati *et al*. huomasivat tutkimuksessaan puheen ja eleiden yhdistämisen sekä nopeuttavan että tehostavan lisätyn todellisuuden ympäristössä toimimista verrattuna pelkkään eleiden käyttämiseen: eleiden avulla käyttäjät pystyivät suoraan vaikuttamaan virtuaalisiin objekteihin, kun taas puhetta käytettiin toimintojen ohjaamiseen erityiskäskyjen avulla.

3.3. Uni- vai multimodaalinen

Vaikka käyttäjät suosivatkin multimodaalista vuorovaikutusta, on tilanteita, joissa he käyttävät pääasiassa unimodaalisia ilmauksia tai sekoittavat keskenään uni- ja multimodaalisia ilmauksia. Siinä missä multimodaalisuus viittaa useamman modaliteetin samanaikaiseen käyttöön, on unimodaalisuus nimensä mukaisesti yhden modaliteetin käyttöä. Syötteiden erottaminen toisistaan joko uni- tai multimodaalisiksi on usein hyvin ongelmallista ja nostaa esille syötteiden tahdistuksen eli synkronoinnin merkityksen.

3.3.1. Syötteiden synkronointi

Unimodaalisiin järjestelmiin verrattuna multimodaaliset järjestelmät tarjoavat luonnollisemman vuorovaikutuksen sekä tehokkaammat toimintamahdollisuudet vähentää virheellisiä syötteitä ja ristiriitaisuuksia. Vastoin yleistä käsitystä käyttäjät eivät kuitenkaan multimodaalisissakaan järjestelmissä anna komentoja jatkuvasti kaikkia mahdollisia modaliteetteja hyödyntäen, vaan sekoittavat keskenään uni- ja multimodaalisia ilmauksia [Oviatt, 1999a].

Oviatt *et al*. [1997] huomasivat tehtävän luonteen selkeästi ennustavan siihen liittyvän komennon antamista multimodaalisesti: esimerkiksi monimutkaisten visuaalisten näyttöjen kanssa työskennellessään käyttäjät todennäköisemmin antoivat avaruudellista sijaintia koskevat käskyt multimodaalisesti kuin unimodaalisesti; toisin kuin käskyt, jotka eivät sisältäneet tilaan tai valintaan liittyvää informaatiota. Käyttäjät toimivat multimodaalisesti 86 %:ssa tehtävistä silloin, kun heidän piti lisätä, liikuttaa, muokata tai laskea kartalla olevien objektin välimatka, jotka kaikki vaativat avaruudellisten suhteiden käsittelyä. Tutkimuksen aikana noin 20 %:a kaikista käskyistä annettiin multimodaalisesti, loput vain yhtä modaliteettia käyttäen.

Erilaisten aikarajojen määrittäminen multimodaalisten syötteiden synkronoinnille on ongelmallista käyttäjien yksilöllisten vuorovaikutustapojen vuoksi. Huang ja Oviatt [Huang and Oviatt, 2006] raportoivat tutkimuksesta, jossa kymmenestä käyttäjästä seitsemän antoivat pääasiallisesti multimodaalisia syötteitä, kun taas kolmella syötteet olivat pääasiassa unimodaalisia. Miltei kaikki käyttäjät antoivat syötteet kahta modaliteettia samanaikaisesti käyttäen. Huolimatta käyttäjien välillä olevasta vaihtelusta, pysyivät yksilöiden omaksumat vuorovaikutustavat yhdenmukaisina kautta tutkimuksen. Kanninen [2003] toteaakin, että käytännössä raja-arvot tulee määrittää tapauskohtaisesti ja

eritoten niissä tilanteissa, joissa käyttäjä joutuu suuntamaan tarkkaavaisuuttaan eri kohteisiin tehtävän aikana.

Synkronoinnin kannalta syötteet voidaan jakaa kahteen syötemalliin (*integration pattern*): samanaikaiseen (*simultaneous*) ja jaksoittaiseen (*sequential*). Aiemman tutkimuksen perusteella noin 70 %:a käyttäjistä antaa syötteet samanaikaisesti ja 30 %:a jaksottaisesti [Huang *et al*., 2006]. Jo tämä kumoaa uskomuksen siitä, että *kaikki* multimodaaliset syötteet annettaisiin samanaikaisesti, jolloin ajallinen päällekkäisyys ratkaisisi sen, mitkä syötteet järjestelmän tulisi yhdistää. Samanaikaisetkaan syötteet eivät ole ajallisesti täysin yhtäaikaisia, vaikka käyttäjät saattavat olettaa näin: multimodaaliset signaalit ovat harvoin täysin samanaikaisia sekä ihmisen ja tietokoneen vuorovaikutuksessa että luonnollisessa kommunikaatiossa [Oviatt, 1999a]. Näennäinen päällekkäisyys ei siis automaattisesti merkitse täyttä samanaikaisuutta.

Gupta [2003a] tutki multimodaalisen navigaatiojärjestelmän avulla käyttäjien syötteiden vaihtelevuutta. Tutkimuksen aikana annetuista syötteistä 83 %:a oli multimodaalisia ja tulos vastaakin näin aiempaa [Oviatt *et al*., 1997] tutkimusta. Käyttäjät antoivat syötteen 95 %:ssa tapauksista kahden modaliteetin avulla, joista eniten käytettiin puhetta ja eleitä. Multimodaalisten syötteiden ajoitus oli samanaikainen 45 %:ssa syötteistä ja lopuissa syöte seurasi toista lyhyen viiveen jälkeen: eri modaliteettien välittämät syötteet erosivat toisistaan keskimäärin 1,5 sekuntia. Gupta myös huomasi unimodaalisten komentojen esittämisen vaativan 18 %:a pidemmän ajan kuin multimodaaliset komennot; tulos, joka tukee multimodaalisuuden tehokkuutta ja nopeutta perinteisiin käyttöliittymiin nähden.

Katseen ja puheen yhdistämistä tutkiessaan Zhang *et al*. [2004] huomasivat, että multimodaalinen järjestelmä ei *kaikissa tilanteissa* suoriutunutkaan paremmin kuin unimodaalinen järjestelmä, vaan virheiden lukumäärä saattoi olla jopa korkeampi kuin yhtä modaliteettia käytettäessä. Huomattava on, että katseen ja puheen synkronointiin vaikuttaa merkitsevästi katseen ennakoiva käyttö: ihmiset katsovat asioita ennen kuin tekevät niille mitään. Katsetta myös siirretään nopeasti, jolloin aiemmasta kohteesta yhä puhuessaan käyttäjä saattaa jo katsoa seuraavaa kohdetta. Kaur *et al*. [2003] analysoivat silmien fiksaatioita (katseen pysähdys) ennen ja jälkeen puheen aloituksen ja huomasivat, että fiksaatio, joka parhaiten määritti kohdeobjektin, tapahtui keskimäärin noin 630 millisekuntia ennen puhetta. Käyttäjien välillä vaihtelu oli jälleen suurta, mutta yksilön tasolla tarkasteltuna suhteellisen pientä.

Multimodaalisten syötteiden synkronointi ei ole yksinkertainen tehtävä. Gupta [2003a] erottelee kaksi käyttäjän kommunikaatiovuoron päättymiseen liittyvää, toisilleen vastakkaista vaatimusta:

- Luonnollisen vuorovaikutuksen saavuttamiseksi ei tulisi käyttää ennalta määrättyjä, vuorovaikutusta rajoittavia vaatimuksia, joissa käyttäjä esimerkiksi pakotetaan antamaan puhekäsky tietyn ajan sisällä näytön koskettamisesta. Huomioitava on, että

jokaisella modaliteetilla on myös oma ajallinen prosessointivaatimus erilaisten resurssien ja taltiointiaikojen takia: esimerkiksi puhesyötteen antaminen kestää kauemmin kuin yksinkertainen osoitusele.

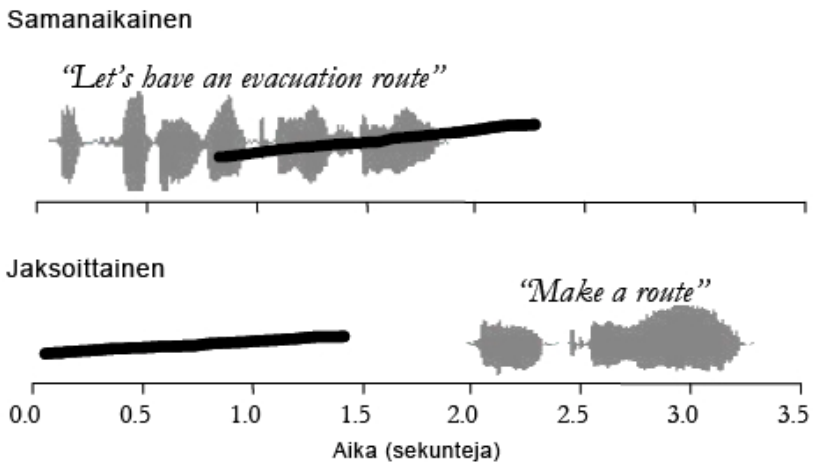
- Käyttäjät odottavat järjestelmän vastaavan välittömästi syötteen antamisen jälkeen. Käyttäjien välisen suuren yksilöllisen vaihtelun takia on kuitenkin hankalaa, ellei mahdotonta, tarkasti määrittää juuri se hetki, jolloin syöte on päättänyt ja järjestelmän annettava vastaus. Rajoitteita käyttämällä olisi mahdollista helpommin erottaa toisistaan uni- ja multimodaaliset syötteet, mutta samalla menetettäisiin multimodaalisten järjestelmien perimmäinen tarkoitus eli ihmisen ja tietokoneen vuorovaikutuksen luonnollisuus.

Yhdessä nämä vaatimukset kiteyttävät multimodaalisiin järjestelmiin läheisesti liittyvän odotusongelman, johon palataan myöhemmin alakohdassa 4.4.1. Käytännössä syötteille on tavallisesti pakko määritellä jonkinlaiset aikarajat, jotka perustuvat useissa tutkimuksissa saatuihin keskivertoaikoihin eri modaliteeteilla annettujen syötteiden päällekkäisyydestä tai viiveestä. Multimodaalisten syötteiden yhdistämistä voidaan myös yrittää välittömästi niiden antamisen jälkeen, jolloin varsinaisia aikarajoitteita ei tarvita. Tämä saattaa kuitenkin helposti johtaa virheellisiin tulkintoihin, jos järjestelmä vahingossa yhdistääkin toisilleen vastakkaiset syötteet.

3.3.2. Samanaikaiset ja jaksoittaiset syötteet

Eri-ikäisten käyttäjien on huomattu valitsevan pääsääntöisesti joko samanaikaisen tai jaksoittaisen syötetavan. Huang *et al*. [2006] summaavat aiempaa tutkimusta toteamalla käyttäjän omaksuman syötetavan olevan miltei välittömästi havaittavissa ja pysyvän pitkälti muuttumattomana sekä tietyssä vuorovaikutustilanteessa (88-97 %:ssa pysyy samana) että ylipäänsä käyttäjän vuorovaikutuksessa tietokoneen kanssa. Lisäksi vallitsevan syötetavan muuttaminen on hankalaa juuri sen pysyvyyden takia. Yksi selittävä tekijä syötetapojen vaihtelevuudelle käyttäjien kesken saattaa olla yksilölliset erot heidän kognitiivisessa tyylissään [Oviatt *et al*., 2005a].

Oviatt *et al*. [2005a] jakoivat tutkimuksessaan käyttäjät samanaikaista tai jaksoittaista syötemallia käyttäviksi, jos vähintään 60 %:a yhden koetilanteen aikana annetuista komennoista ilmaistiin käyttäen toista näistä malleista. Jos lukumäärä jäi tämän alle tai käyttäjä yllättäen vaihtoi syötemallia, ei syötemallia pidetty hallitsevana. Tutkimuksen aikana huomattiin, että käyttäjät pitäytyivät omaksumassaan syötemallissa läpi tutkimuksen ja vain kahdesti ilmeni tilanne, jossa käyttäjät väliaikaisesti jäivät alle 60 %:n rajan. Modaliteettien käytössä huomattiin eroavaisuuksia siinä, aloitettiinko puhe- vai kynäsyötteellä ja mikä oli niiden ajallinen suhde toisiinsa. Samanaikaista syötetapaa käyttävillä puhe edelsi kynäsyötettä ja modaliteettien keskimääräinen ajallinen päällekkäisyys oli 1 sekunti; jaksoittaista syötetapaa käyttävillä kynäsyöte edelsi puhetta ja keskimääräinen ajallinen viive syötteiden välillä oli 0,6 sekuntia (Kuva 1).



Kuva 1. Samanaikaisen ja jaksoittaisen syötemallin eroavuus [mukailtu Oviatt *et al*., 2005a].

Samansuuntaisia tuloksia saavuttivat myös Oviatt *et al*. [1997], joskin heidän tutkimuksessaan samanaikaista syötemallia käytettäessä kynäsyöte edelsi puhetta 57 %:ssa tapauksista ja vain 14 %:ssa annettiin puhesyöte ennen kynää. Jaksoittaista tapaa käytettäessä kynäsyöte tehtiin 99 %:ssa tapauksista valmiiksi ennen puheen aloittamista ja keskimääräinen viive näiden välillä oli 1,4 sekuntia. Yksi syy kynäsyötteen ensisijaisuuteen saattoi olla käyttäjien tarve käsitellä ja muokata informaatiota mielessään ennen sen kielellistä esittämistä. Oviatt *et al*. luokittelivat tutkimuksessaan syötetavat neljään kategoriaan:

1. Samanaikainen (simultaneous), jossa puhe ja kynäsyöte annetaan samanaikaisesti.
2. Jaksoittainen (sequential), jossa jompikumpi – puhe tai kirjoitus – edeltää toista modaliteettia, joka seuraa perässä viiveellä.
3. Ele & puhe (point & speak), jossa käyttäjä osoittaa objektia samalla siitä puhuen. Muita piirtomerkkejä ei kuitenkaan tehdä yksittäisen pisteen lisäksi.
4. Yhdistelty (compound), joka jakautuu kahteen osaan sisältäen kirjoitusvaiheen ja ele- & puhevaiheen.

Näistä syötemalleista samanaikaista käytettiin 42 %:ssa tehtävistä, jaksoittaista 32 %:ssa, elettä & puhetta 14 %:ssa ja yhdisteltyä vain 12 %:ssa.

Tutkimustulokset korostavat syötteiden ajoituksen merkitystä multimodaalisessa vuorovaikutuksessa. Oviatt *et al*. [2003] huomasivat syötemallin pysyvyyden osalta käyttäjien taipumuksen ”venyttää” omaksumaansa mallia vuorovaikutustilanteen mukaan: esimerkiksi virhetilanteessa samanaikaista syötemallia käyttävillä modaliteettien ajallinen päällekkäisyys nousi 1,5 sekunnista 1,77 sekuntiin. Samalla tavalla jaksoittaista mallia käyttävillä viiveen ajallinen kesto kasvoi. Sama muutos huomattiin myös tehtävien vaikeutuessa, jolloin sekä modaliteettien jaksoittainen viive että samanaikainen

päällekkäisyys kasvoivat tasaisesti vaikeustason noustessa. Oviatt *et al*. [2003] arvioivat syötemallien korkean pysyvyyden johtuvan osin siitä, että käyttäjät ovat omaksuneet hyväksi havaitun tavan (*success strategy*) olla vuorovaikutuksessa tietokoneiden kanssa ja tahtovat pysyä siinä välttääkseen vaikeudet.

3.4. Vuorovaikutustapa

Aiemmin on jo useasti viitattu käyttäjien omaksumaan yksilölliseen tapaan toimia vuorovaikutustilanteissa. Huang ja Oviatt [Huang and Oviatt, 2006] huomasivat käyttäjän vuorovaikutustavan olevan tunnistettavissa 100 %:n varmuudella 15 ensimmäisen komennon jälkeen. Joidenkin käyttäjien omaksuma tapa oli nähtävissä jo viiden komennon perusteella, mutta luotettavin tulos saatiin useampia komentoja tarkastelemalla. Käyttäjien hallitsevan vuorovaikutustavan tietäminen voi selittää 85 %:a varianssista, joka liittyy käyttäjien todennäköisyyteen toimia multimodaalisesti seuraavan syötteen yhteydessä [Huang *et al*., 2006]. Yksilöllisyyden takia on erilaisten vuorovaikutusta rajoittavien aikakynnysten tai kielellisen rajoitteiden asettaminen vaikeaa. Ne lisäksi rikkoisivat multimodaalisen vuorovaikutuksen luonnollisuuden, kun käyttäjät joutuisivat mukauttamaan toimintansa järjestelmän vaatimuksiin.

Selityksen yksilöiden väliselle vaihtelulle vuorovaikutustavoissa saattaa tuoda kognitiivinen tyyli, joka on yksilön luonteenomainen tapa reagoida ympäristöön [Oviatt *et al*., 2005a]. Kognitiivinen tyyli vaikuttaa siihen, miten ihmiset havaitsevat, ajattelevat ja muistavat informaatiota. Tutkijat ovat käyttäneet useita erilaisia nimityksiä tyyleille, joista yksi on jako reflektiivisyyteen ja impulsiivisuuteen. Reflektiivinen ihminen tekee tarkkoja ja yksityiskohtaisia havaintoja sekä toimii harkitusti, kun taas impulsiivinen ihminen toimii nopeasti, jolloin havainnot helposti jäävät epätarkoiksi ja hajanaisiksi [Riding and Cheema, 1991]. Vaikka nämä erot reflektiivisessä ja impulsiivisessa tyylissä ovat suhteellisen pysyviä, voidaan niihin jossain määrin vaikuttaa kokemusten ja harjoittelun avulla [Oviatt *et al*., 2005a].

Yksilöllisiä eroja tutkiessaan Oviatt *et al*. [2005a] huomasivatkin samanaikaista ja jaksoittaista syötemallia käyttävien välillä selviä eroja sekä käyttäytymisessä että kognitiivisessa tyylissä. Jaksoittaisen syötemallin käyttäjät tekivät merkittävästi vähemmän tehtävien kannalta kriittisiä virheitä, varsinkin juuri esitellyissä ja monimutkaisissa tehtävissä. Odotusten vastaisesti he eivät kuitenkaan olleet hitaampia kuin samanaikaisen syötemallin käyttäjät. Jaksoittaisen mallin omaksuneet käyttivät myös lyhyempiä kielellisiä käskyjä, kun taas samanaikaisen mallin käyttäjien lauseet olivat pidempiä, epäsuoria ja keskusteluun pyrkiviä (sanasto laajempaa). Oviatt *et al*. [2005a] tulkitsivat jaksoittaisen syötemallin käyttäjien pyrkivän tiiviin sanaston avulla varmistamaan virheettömän ja kommunikaation onnistumiseen tähtäävän vuorovaikutuksen. Jaksoittainen syötemalli voidaan näin ollen yhdistää yksilöissä reflektiiviseen kognitiiviseen tyylin ja samanaikainen malli impulsiivisuuteen.

Kognitiivisen tyylin lisäksi vuorovaikutukseen vaikuttavat annetut ohjeet ja kohdealueen asiantuntijuus sekä yksilön fysiologinen tila. Oviatt *et al*. [2005b] tutkivat voidaanko ihmisiä kannustaa vaihtamaan multimodaalista vuorovaikutustapaansa selkeiden ohjeiden avulla ja pysyykö vaihtunut tapa muuttumattomana kuukauden mittaisen seurantatutkimuksen aikana. Tutkimuksessa vain 37 %:a käyttäjistä vaihtoi vuorovaikutustapaansa ja pysyi siinä seurannan aikana. Käyttäjistä 19 %:a ei muuttanut tapaansa ollenkaan ja 31 %:a vaihtoi, mutta palaisi seurantatutkimuksen aikana takaisin alkuperäiseen vuorovaikutustapaan. Yhtenä ohjeiden tehottomuutta selittävänä tekijänä tutkijat pitivät käyttäjien rajoittunutta tietoisuutta omasta multimodaalisesta vuorovaikutustavastaan: esimerkiksi 62 %:a käyttäjistä ei muistanut tai muisti väärin sen oliko heidän omaksuman tapa muuttunut tutkimusjaksojen välillä.

De Angeli *et al*. [1998] huomasivat tietokoneen käytön hallitsevien käyttäjien suosivan multimodaalisia syötteitä verrattuna aloittelijoihin, joilla taasen ei ollut selvää mieltymystä uni- ja multimodaalisten syötteiden välillä. Myös Kaur *et al*. [2003] havaitsivat käyttäjien harjaantuneisuustason vaikutukset: virheiden määrä oli selkeästi suurin aloittelevalla käyttäjällä. Vuorovaikutukseen vaikutti myös käyttäjien väsymys, joka ilmeni virheiden lukumäärän nousemisella tutkimuksen myöhemmissä vaiheissa. Samalla tavalla vaikutusta voidaan ajatella olevan käyttäjän muilla fysiologisilla tiloilla (nälkä, jano) tai psyykkisellä tilalla: esimerkiksi masennuksen on todettu vaikuttavan ihmisten toiminnan nopeuteen ja varmuuteen.

Tulevaisuudessa multimodaalisten järjestelmien onkin kyettävä sopeutumaan käyttäjien yksilöllisiin vuorovaikutustapoihin eikä pyrkiä ohjaamaan käyttäjiä tietyn mallin ja rajoitusten alle. Huang ja Oviatt [Huang and Oviatt, 2006] laskevat käyttäjien toimintaan mukautuvien aikarajojen sekä vähentävän multimodaalisen järjestelmän prosessointiviivettä noin 50-60 %:a että parantavan multimodaalista tulkintaa noin 50 %:a. Lisäksi käyttäjien ja järjestelmän välisen vuorovaikutuksen synkronisaatio ja yleensä suorituskyky paranisivat. Virheiden estämiseksi ja käytön miellyttävyyden parantamiseksi multimodaalisten järjestelmien tulisi muun muassa tukea tarkkaavaisuuden ylläpitoa ja impulsiivisen tyylin omaavia käyttäjiä, joilla virheherkkyys on suuri [Oviatt *et al*., 2005a]. Kognitiivisen tyylin voimakas vaikutus vuorovaikutukseen osoittaakin, että multimodaalisten järjestelmien suunnittelussa ja kehityksessä tulee tulevaisuudessa hyödyntää monitieteistä näkökulmaa pelkän tietojenkäsittelytieteen sijaan.

4. Multimodaalisten syötteiden integraatio

Kun eri modaliteettien kautta tulleet syötteet on kerätty, on ne yhdistettävä yhden yksiselitteisen merkityksen omaavaksi syötteeksi. Vasta tämän perusteella tietokone voi suorittaa käyttäjän antaman komennon. Vaikeaa syötteiden yhdistämisestä yhdenmukaiseksi tekee multimodaalisen vuorovaikutustavan suuri vaihtelevuus eri käyttäjien välillä. Myös käytetty tekniikka ja sovellus itsessään voivat vaikeuttaa yhdistämisen eli fuusion onnistumista. Corradini *et al*. [2003] huomauttavat aiemman kokeellisen tiedon merkityksestä, jonka analysointi voi auttaa valitsemaan sopivimman lähestymistavan sovellukseen, modaliteetteihin, käyttäjiin ja tehtävään nähden.

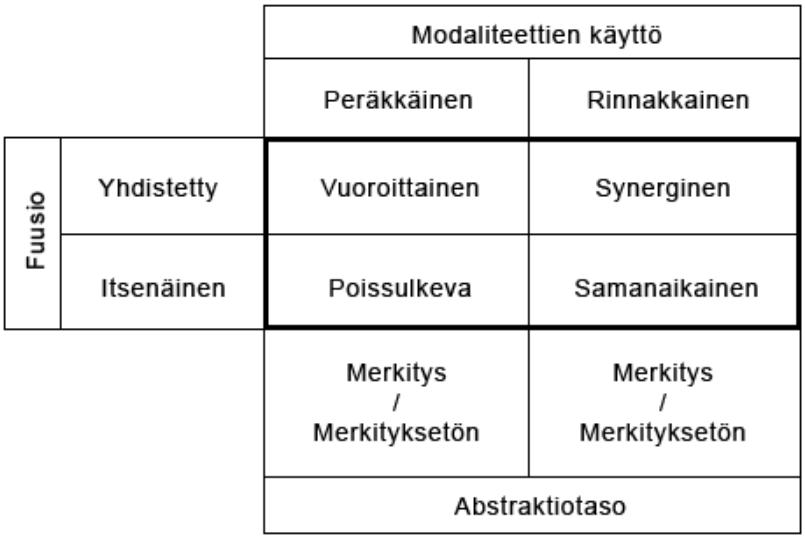
4.1. Fuusioprosessi

Fuusioprosessia on luonnehdittu useiden kriteerien avulla. Salber *et al*. [1995] kuvaavat sitä kahdella määritelmällä:

a) Eri informaatiotyyppien abstrahointi/konkretisointi uuteen muotoon toiselle prosessille, kun erilaista informaatiota vastaanotetaan erillisiltä prosesseilta.

b) Useamman informaatiotyypin yhdistäminen jollakin abstraktiotasolla yhdeksi informaatiotyypiksi, joka on samalla abstraktiotasolla.

Nämä yleisluontoiset määritelmät kuvaavat laajasti kaikki fuusioprosessit. Lisäksi fuusiota voidaan tarkistella yhdistettävien syötteiden ajallisina suhteina, kuten Nigay ja Coutaz [Nigay and Coutaz, 1993] tekivät. Heidän ehdottamansa luokittelu jakaa modaliteettien käyttötavat kolmen ulottuvuuden mukaan, joita ovat abstraktiotaso, modaliteettien käyttö ja fuusio (Kuva 2). Abstraktiotasoltaan multimodaaliset järjestelmät kuuluvat merkitys-kategoriaan edellyttäen, että syötteiden merkitys on tiedossa (esimerkiksi puhesyöte voi olla joko pelkkä signaali tai valmis lause) ja vuorovaikutus itsessään on tulkittavissa merkitykselliseksi. Modaliteettien käyttö viittaa siihen, ovatko eri modaliteetit käytettävissä peräkkäin yksi kerrallaan (*sequential*) vai rinnakkaisesti (*parallel*). Fuusio-ulottuvuus vuorostaan jaottelee vuorovaikutuksen sen mukaan, onko syötteet yhdistettävä keskenään (*combined*) vai voidaanko niitä käsitellä itsenäisinä (*independent*).



Kuva 2. Luokittelu multimodaalisille järjestelmille [mukailtu Nigay and Coutaz, 1993].

Jos syötteet tulee yhdistää keskenään ja modaliteettien käyttö on peräkkäistä, on vuorovaikutus vuoroittaista (*alternate*). Modaliteettien ollessa samanaikaisesti käytössä on vuorovaikutus synergista (*synergistic*). Nämä kaksi kategoriaa määrittävät syötteiden väliset ajalliset suhteet. Nigayn ja Coutazin luokittelusta voidaan tehdä myös muita huomioita. Ensinnäkin jokaisessa luokittelun kategoriassa syötteiden käsittely poikkeaa muista. Yksinkertaisimmillaan syöte voi olla poissulkeva (*exclusive*), jolloin vuorovaikutus on unimodaalista ilman tarvetta yhdistää käyttäjän antamia syötteitä keskenään. Suurimmat toteutushaasteet ovat synergia-kategoriassa, joka toisaalta tarjoaa myös tehokkaimman vuorovaikutuksen ihmisen ja tietokoneen välille. Vuoroittaiset syötteet puolestaan tuovat mukanaan odotusongelman, johon palataan tarkemmin alakohdassa 4.4.1.

Vo ja Waibel [Vo and Waibel, 1997] ottavat esille kaksi kysymystä, joihin syötteiden fuusiossa tulisi saada vastaukset:

- Mikä merkitys multimodaaliselle syötetapahtumalle (*input event*) kokonaisuudessaan annetaan?
- Miten tämä merkitys saadaan, kun yhdistetään eri modaliteettien kautta tulleet syötteet?

Suorittaakseen oikean toiminnon vastauksena käyttäjän antamaan syötteeseen, on tietokoneen pystyttävä tulkitsemaan käyttäjän toimet oikein ja näin johtamaan niistä yksiselitteinen komento, jonka perusteella voidaan antaa vastaava tuloste. Jotta tulkintaprosessi voidaan suorittaa ongelmitta, tulee järjestelmällä olla riittävät mahdollisuudet yhdistää yhdestä tai useammasta lähteestä tulevat syötesignaalit keskenään. Käytännössä tämä tarkoittaa jokaisen käytettävissä olevan modaliteetin kautta tulleen syötteen tallentamista ja sopivaa toimintamallia ratkaisemaan kysymys siitä, tuleeko erilliset syötteet yhdistää vai ei.

Todellisuudessa suunnittelijoiden on teoreettisten pohdintojen sijaan valittava yksi selkeästi määritelty fuusiomenetelmä, jonka toiminta on johdonmukaista kautta sovelluksen ja lisäksi kiinnitettävä huomiota niihin käyttöliittymän ominaisuuksiin, joihin fuusioprosessi itsessään vaikuttaa [Faconti *et al*., 1996]. Multimodaaliselle järjestelmälle sopivan ja tehokkaan fuusiomenetelmän valitseminen on kriittistä, sillä ilman sitä käyttäjän antaman komennon tulkinta voi jäädä osittain puutteelliseksi tai tulla kokonaan väärinymmärretyksi. Fuusioprosessin onnistumisen varmistaminen onkin yksi niistä haasteista, joita multimodaalisuus tuo käyttöliittymien suunnittelijoille.

4.2. Kolme tasoa

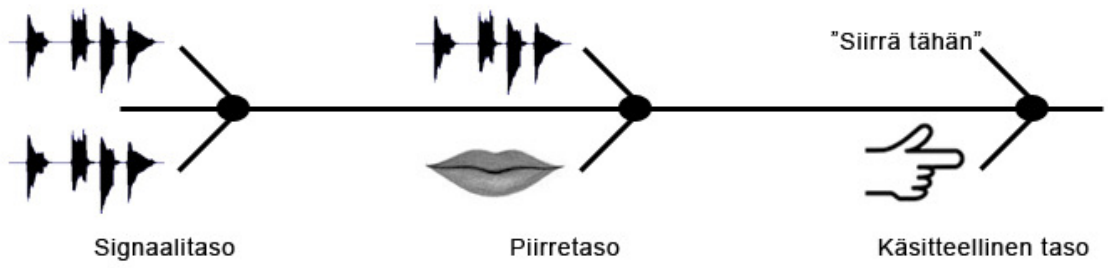
Yleisesti ottaen syötteiden fuusioprosessi voidaan suorittaa kolmella eri tasolla [Vo and Waibel, 1997; Paleari and Lisetti, 2006; Russ *et al*., 2005]. Alimmalla eli signaalitasolla (*signal/data level*) fuusio tarkoittaa kahden tai useamman tyypiltään samanlaisen signaalin yhdistämistä: molemmat ovat esimerkiksi äänisignaaleja. Tällä tasolla voidaan yhdistää vain hyvin läheiset ja keskenään synkroniset signaalit, sillä informaatio on huomattavasti yksityiskohtaisempaa kuin myöhemmissä vaiheissa. Syötteiden fuusion suorittaminen näin varhaisessa vaiheessa on tyypillisesti hankalaa eikä edes multimodaalisten järjestelmien kohdalla kannattavaa, kun otetaan huomioon modaliteettien vaatimat erilaiset vuorovaikutuslaitteet, jolloin myös saadut syötteet poikkeavat toisistaan.

Keski- eli piirretasolla (*feature level*) käsittelemättömät syötesignaalit muutetaan helpommin ymmärrettävään muotoon ennen fuusiota ja tulkintaa: esimerkiksi äänisignaali muutetaan merkkijonoksi. Fuusioprosessissa eri syötteiden piirteet yhdistetään keskenään, mikä edellyttää syötteiltä riittävää synkronisaatiota, jotta tulkinnan tulos olisi tyydyttävä. Piirretason fuusio on käyttökelpoinen multimodaalisissa järjestelmissä, mutta vaatii tietokoneelta paljon laskennallista tehoa piirteiden määrän ja ominaisuuksien vaihdellessa. Tällä tasolla voidaan yhdistää keskenään esimerkiksi puhe ja huulten liike sekä puhe ja videokuva tunnetilaa tunnistettaessa.

Ylimmällä eli käsitteellisellä tasolla (*decision/conceptual level*) fuusio perustuu yksittäisen modaliteetin syötteen tulkitsemiseen: jokaiselle syötteelle annetaan oma osittainen tulkinta ja lopuksi nämä puolittaiset tulkinnat yhdistetään yhdeksi tulkinnaksi. Hyväksi käytetään syötteestä suoraan johdettua merkitystä: käden liike tulkitaan osoituseleeksi ja puhe tunnistetaan tietyksi komennoksi. Tällä tasolla syötteiden synkronisoinnin merkitys ei ole niin suuri kuin kahdella aikaisemmalla ja samalla fuusioprosessi voidaan toteuttaa yksinkertaisten algoritmien avulla. Multimodaalisissa järjestelmissä syötteiden fuusio suoritetaan tavallisesti käsitteellisellä tasolla.

Fuusioprosessia eri tasoilla havainnollistaa kuva 3. Signaalitasolla kaksi samanlaista, mutta eri lähteistä (esimerkiksi äänisignaalien yhteydessä käytetään vähintään kahta mikrofonia) tulevaa signaalia voidaan yhdistää paremman signaalin saavuttamiseksi. Parannellusta signaalista voidaan näin piirretasolla helpommin erotella erilaisia piirteitä, jotka yhdistetään toisesta signaalista eroteltujen piirteiden kanssa. Tämä helpottaa tul-

kintaa, kun toisiinsa liittyvät syötteet voidaan yhdistää (huulten liikkeiden yhdistäminen puhesyötteeseen parantaa entisestään tunnistamista). Käsitteellisellä tasolla voidaan yhdistää kaksi täysin erityyppistä syötettä: kuvassa aiemmissa vaiheissa käsitelty puhesyöte yhdistetään osoituseleen kanssa, jotta saadaan selville, mikä käyttäjän toiminnan tavoite oli.



Kuva 3. Syötteiden fuusio voidaan suorittaa kolmella erillisellä tasolla.

Käsitteellisen tason edut muiden tasojen fuusioon verrattuna ovat huomattavat. Corradini *et al*. [2003] muistuttavat tällä tasolla jokaisen modaliteetin vuorovaikutuslaitteiden olevan erillisiä, jolloin ne voidaan mukauttaa (*training*) erikseen ja yhdistää ilman uudelleen mukauttamista, kun taas piirretasolla siihen vaaditaan suuret tietomäärät. Lisäksi ylimmällä tasolla voidaan hyödyntää kaupallisia laitteita perusmodaliteeteille, kuten puheelle. Avainsanana on yksinkertaisuus: syötteiden yhdistäminen ei vaadi ylimääräisiä parametreja niiden lisäksi, joita modaliteettien vuorovaikutuslaitteet käyttävät. Tämä mahdollistaa yleistämisen modaliteettien määrän ja tyypin perusteella. Toisaalta käsitteellisen tason fuusion edellytyksenä on multimodaalisten syötteiden esiprosessointi alimmilla tasoilla: esimerkiksi äänisignaali on jo muutettu sanoiksi tai osoitusele koordinaateiksi.

4.3. Arkkitehtuurit ja agentit

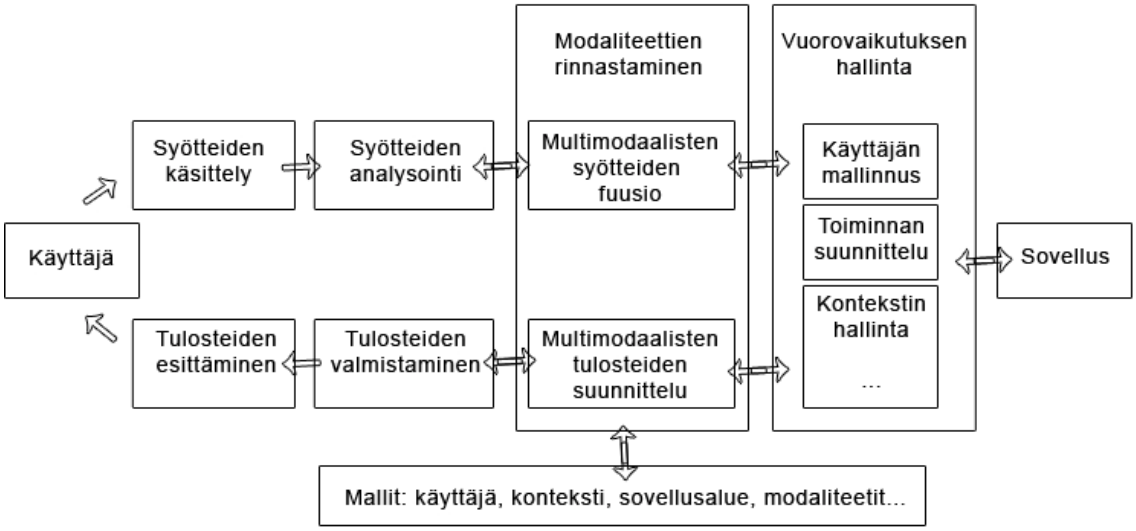
Multimodaalista fuusioprosessia tarkasteltaessa ei pidä unohtaa arkkitehtuurin merkitystä, sillä se toimii suunnittelun ja toteutuksen pohjana. Useimmissa multimodaalisissa arkkitehtuureissa on oma tasonsa syötteiden yhdistämiselle, mutta yksimielisyyteen yleisluontoisesta näkemyksestä ei ole päästy. Toinen fuusion kannalta tärkeässä osassa oleva tekijä on agentit. Useat järjestelmät hyödyntävät niitä syötteiden käsittelyn eri vaiheissa. Koska nämä molemmat aihepiirit ovat laajat ja niiden läpikäyminen kokonaisuudessaan ei ole tämän tutkielman kannalta tarkoituksenmukaista, käsitellään arkkitehtuureja ja agentteja tässä vain yleisellä tasolla.

Kehitettäessä arkkitehtuureja multimodaalisille järjestelmille tulee ottaa huomioon alueen monimuotoisuus ja laaja-alaisuus. Bunt *et al*. [2003] listaavat muutamia toiminnallisia vaatimuksia:

- tuki modaliteettien integraatiolle (syötteiden fuusio ja soveltuva tuloste),

- tilanteeseen (käyttäjä, tehtävä, sovellus) sopiva reaaliaikainen havaitseminen ja palaute,
- tuki lisääntyvälle prosessointitarpeelle ja jatkokehitykselle, ja
- skaalautuvuus.

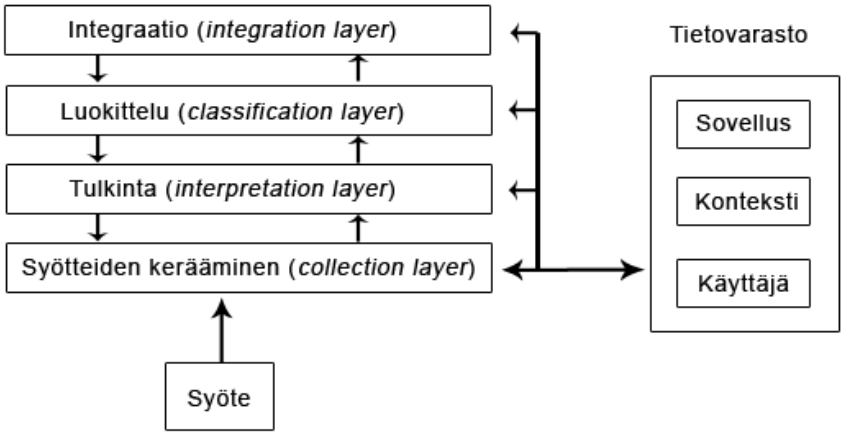
Näiden lisäksi on myös useita keskeisiä järjestelmään ja tekniikkaan liittyviä vaatimuksia, joita multimodaalisten järjestelmien tulisi tukea. Erilaisia arkkitehtuureja analysoituaan Bunt *et al*. muokkasivat Mayburyn ja Wahlsterin [Maybury and Wahlster, 1998] esittämää korkean tason multimodaalista arkkitehtuuria, joka sisältää kaikki vaiheet käyttäjän syötteestä tietokoneen tulosteeseen (Kuva 4). Syötteiden fuusiota edeltää prosessointi- ja analyysivaihe, jonka jälkeen suoritetaan fuusio ja tulkinta, jota muodostettaessa käytetään hyväksi vuorovaikutuksen sen hetkistä tilaa, kontekstia (aika, paikka, tehtävä) ja tietoja käyttäjästä. Kokonaisuudessaan arkkitehtuuri tarjoaa abstraktin näkemyksen multimodaalisen käyttöliittymän toteutuksesta.



Kuva 4. Korkean tason multimodaalinen arkkitehtuuri [mukailtu Bunt *et al*., 2003].

Konkreettisemman arkkitehtuurin multimodaalisten syötteiden käsittelyyn tarjoaa Gupta [2003b]. Arkkitehtuuri koostuu tietovarastosta ja tasoista, joilla on omat tehtävät, prosessit ja järjestelmäkomponentit (Kuva 5). Tietovarasto on kaikkien tasojen käytettävissä oleva kokoelma, joka sisältää tietoja sovelluksesta (tila- ja tehtävämalli), kontekstista (parhaillaan suoritettava tehtävä, vuorovaikutuksen historia, ympäristö) ja käyttäjästä (omaksuttu vuorovaikutustapa, tiettyjen modaliteettien suosiminen). Alin taso vastaa käsittelemättömien syötteiden keräämisestä, jonka jälkeen ne tulkintatasolla muutetaan modaliteetista riippumatta syntaktisiksi rakenteiksi tai jopa semanttisiksi tulkinnoiksi. Luokittelutasolla useiden modaliteettien kautta tulleiden syötteiden semanttiset tulkinnat liitetään yhteen käyttämällä sekä ajallisia että merkitykseen liittyviä yhdistelysääntöjä. Lopuksi integraatiotasolla kaikki käyttäjän vuoron aikana kerätyt tulkinnat fuusioidaan keskenään yhdeksi tai useammaksi yhteistulkinnaksi, jotka sisältävät

informaatiota kaikista yksittäisistä tulkinnoista. Nämä siirtyvät järjestelmässä eteenpäin tulosteesta vastaavalle alajärjestelmälle.



Kuva 5. Viitemalli multimodaalisten syötteiden tulkinnalle [mukailtu Gupta, 2003b].

Multimodaaliset järjestelmät toteutetaan usein agentteja käyttäen. Agentti on yleisluontoinen määritelmä tietokoneohjelmalle, joka pystyy suorittamaan määrättyjä tehtäviä itsenäisesti tai yhteistyössä muiden kanssa. Monimutkaisten järjestelmien toteuttaminen niiden avulla onkin helpompaa, kun tehtävät voidaan jakaa erikoistuneiden agenttien kesken. Lisäksi useiden agenttien yhteistyötä hyödyntämällä voidaan yksittäisen agentin toteutus pitää yksinkertaisena, vaikka hallittavissa olisi vaikeasti käsiteltävä tehtäväkokonaisuus.

Yksi tapa toteuttaa multimodaalinen järjestelmä agenttien avulla on käyttää hyväksi moniagenttiarkkitehtuureja, joihin kuuluu esimerkiksi OAA (*Open Agent Architecture*). Se tukee useiden modaliteettien mukanaoloa, mahdollistaa agenttien lisäämisen jo valmiiseen järjestelmään ja soveltuu käytettäväksi mobiililaitteiden kanssa [Moran *et al*., 1997]. Agentit voidaan määrätä huolehtimaan kaikista syötteiden käsittelyyn kuuluvista vaiheista: ne voivat vastaanottaa eri modaliteeteilta tulevat syötteet, käsitellä ja tulkita ne sekä hoitaa fuusioprosessin. Moran *et al*. [1997] käyttävät fuusion toteuttamiseen erikoistunutta agenttia (*modality coordination agent*), jonka tehtäviin kuuluu myös viittaussuhteiden selvittäminen, puuttuvien tietojen täydentäminen ja ristiriitaisuuksien ratkaiseminen käyttäen hyväksi kontekstia, syötteiden vastaavuuksia tai niiden tarpeettomuutta.

4.4. Ongelmia

Multimodaalisilla järjestelmillä on oikein toteutettuna parempi toimintavarmuus verrattaessa yksittäisiin syötteiden tunnistamistekniikoihin, jotka jo luonnostaan ovat virheherkkiä. Oviatt [1999a] huomauttaa tavallisena uskomuksena kuitenkin olevan, että kahden virheherkän tunnistamistekniikan yhdistäminen johtaa kaksinkertaiseen virhemäärään ja suurempaan tunnistamisen epävarmuuteen. Käytännössä multimodaalisuus

tekee virheiden käsittelyn ongelmista helpompia ratkaista, mikä ei kuitenkaan tarkoita sitä, että syötteisiin liittyvät ongelmat olisivat yhdentekeviä. Esimerkiksi puheentunnistuksessa tunnistamistarkkuus huononee aina, kun käyttäjän puhetyyli jollakin tavalla poikkeaa siitä harjoitusaineistosta, jolla tunnistin kehitettiin [Oviatt, 2000]. Yleisesti ottaen tunnistettavien syötteiden kanssa toimittaessa on loppujen lopuksi mahdotonta täysin välttyä ristiriitaisuuksilta.

Yksi multimodaalisuuden keskeisimmistä ongelmista on niin sanottu odotusongelma, jota unimodaalisissa järjestelmissä ei ole. Odotusongelma nimensä mukaisesti viittaa siihen, kuinka kauan järjestelmän on odotettava ennen kuin syöte voidaan tulkita uni- tai multimodaaliseksi. Muita haasteita ovat muun muassa ristiriitaiset, epätarkat ja lyhennellyt syötteet sekä sopimaton tai liian monimutkainen modaliteettien viittaussuhteiden käyttö [Chai, 2002].

4.4.1. Odotusongelma

Odotusongelma liittyy läheisesti alakohdassa 3.3.1. käsiteltyyn syötteiden synkronointiin eli siihen tuleeko syötteitä käsitellä uni- vai multimodaalisina. Nigayn ja Coutazin [Nigay and Coutaz, 1993] mallissa odottaminen on pakollista vuoroittaisten syötteiden kohdalla, jolloin järjestelmän on pystyttävä erottamaan peräkkäisistä syötteistä yhteen kuuluvat kokonaisuudet. Kaikki tämä on vahvasti kytköksissä käyttäjän omaksumaan multimodaaliseen vuorovaikutustapaan: miten he yhdistelevät modaliteetteja, pysyykö omaksuttu tapa yhdenmukaisena kautta tehtävien ja millainen käyttäjän kognitiivinen tyyli on.

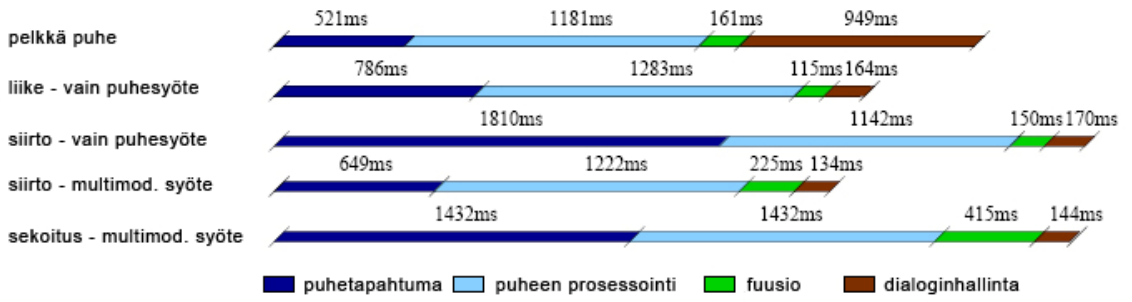
Yhteenvetona voitaisiin tiivistää seuraavasti: multimodaalisen järjestelmän on sekä pääteltävä, koska syöte alkaa ja koska se loppuu, että tarvittaessa yhdistettävä ajallisesti lähekkäin annetut syötteet. Streit [2001] muistuttaa odottamisen olevan tässä välttämätöntä, sillä myöhempi syöte voi ilmaista tavoitteen, joka ensimmäisen syötteen kanssa johtaisi täysin erilaiseen toimintoon kuin yksinään käsiteltynä. Lisäksi koko prosessin tulee tapahtua riittävän nopeasti, jotta käyttäjä ei turhaudu järjestelmän toimintaan. Odotusongelmassa onkin kyse juuri nopeasta toiminnasta: kuinka kauan voi odottaa ilman, että viive häiritsee vuorovaikutusta käyttäjän kanssa?

Gupta [2003a] esittää neljä seikkaa, jotka auttavat uni- ja multimodaalisen syötteen erottelussa ja joita voidaan hyödyntää suunniteltaessa mukautuvaa odotusmekanismia:

- Jos modaliteetti on erikoistunut eli sitä käytetään tavallisesti unimodaalisena, on todennäköisyys saada informaatiota toisen modaliteetin kautta hyvin vähäinen.
- Jos modaliteettia tavallisesti käytetään yhdessä muiden modaliteettien kanssa, on todennäköisyys saada informaatiota toisen modaliteetin kautta suuri.
- Jos informaatio on jakautunut useampaan kuin kahteen tai kolmeen osaan vuorovaikutusvuoron aikana, on todennäköisyys saada lisäinformaatiota muiden modaliteettien kautta vähäinen.

- Jos syötteen kesto yksittäisen modaliteetin kautta on suurempi kuin tavallisesti, on todennäköistä, että käyttäjä on antanut kaiken tiedon unimodaalisesti kyseisellä modaliteetilla.

Ajallista vaihtelua uni- ja multimodaalisten syötteiden kesken havainnollistaa kuva 6. Flippo *et al*. [2003] mittasivat vasteajat viidelle erilaiselle puhesyötteelle, jotka esiintyivät joko yksinään tai osana multimodaalista syötettä (puhe ja kosketus). Vasteajoissa suurimmat erot ovat itse puhetapahtuman (*speech act*) aikana riippuen siitä oletetaanko kyseessä olevan uni- vai multimodaalinen syöte. Nämä erot johtuvat juuri vaatimuksesta odottaa ennalta määrätty aika ennen kuin voidaan otaksua käyttäjän päättäneen syötteen antamisen. Flippo *et al*. huomauttavat myös, ettei näihin viiveisiin voida niiden luonteen takia vaikuttaa nopeammalla laitteistolla, vaan odotusongelmaa on lähestyttävä muilla keinoilla: esimerkiksi hyödyntämällä tilastotietoja käyttäjien vuorovaikutustavoista.



Kuva 6. Puhesyötteen vasteajat viidelle syötetapahtumalle [mukailtu Flippo *et al*., 2003].

Yksi ratkaisu odotusongelmaan on eri modaliteeteilla annettujen syötteiden järjestyksen ja niiden välisen keskimääräisen viiveen perusteella ennustaa todennäköisyyttä sille, onko syöte uni- vai multimodaalinen [Oviatt *et al*., 1997]. Oviatt *et al*. [2005a] huomauttavat nykyisten multimodaalisten järjestelmien käyttävän kiinteitä aikakynnyksiä (*fixed temporal thresholds*), jotka pohjautuvat aikaisempaan malliin käyttäjien luonnollisesta modaliteettien yhdistämisestä. Paremmin käytettäväksi sopisivat kuitenkin käyttäjään mukautuvat aikakynnykset (*user-adaptive temporal thresholds*), jotka sopeutuvat käyttäjien vaihteleviin multimodaalisiin vuorovaikutustapoihin. Ne parantaisivat järjestelmän prosessointinopeutta vähentämällä viivettä 44 %:a nykyisestä sekä lisäisivät luotettavuutta ja vuorovaikutuksen synkronisaatiota. Toisin sanoen mukautuvat aikakynnykset voivat vähentää odotusaikaa, kun järjestelmälle annetaan mahdollisuus toimia joustavasti käyttäjän mukaan.

Streit [2001] pyrkii ratkaisemaan odotusongelman tukemalla multimodaalisten ilmausten synkronisaatiota, jotta syötteiden välit pysyisivät lyhyinä ja tunnistamalla ne tapaukset, joissa välit ovat merkityksettömiä (yhdistämistä ei tarvita). Synkronisaation tukeminen onnistuu esimerkiksi graafisista käyttöliittymistä tutulla tavalla: kohteen va-

lintaan ei riitä pelkkä hiiren osoittimen kuljettaminen sen yli, vaan vaaditaan myös tarkoituksenmukainen komento. Palautteen (kohteen korostus) avulla käyttäjä ymmärtää, että eleiden lisäksi on annettava puhekomento, jotta elettä voidaan pitää tavoitteellisena kommunikaationa. Lisäksi Streit huomauttaa, että jos eleistä voidaan osoittaa niiden olevan joko ainoastaan lisäys viittaukseen tai muokkaava toimenpide, voisi tämä ratkaista odotusongelman jälkimmäisessä tapauksessa.

4.4.2. Syötteiden tunnistaminen

Odotusongelman lisäksi multimodaalisille järjestelmille hankaluuksia tuottaa syötteiden laatu ja niiden viittaussuhteet. Syötteet voivat olla merkitykseltään ristiriitaisia tai epätarkkoja ja niiden tunnistaminen voi epäonnistua. Ne eivät välttämättä myöskään tarjoa riittävästi tietoa tulkinnan tekemiseen: käyttäjillä on tapana antaa uutta tietoa vasta, kun on heidän vuorovaikutusvuoronsa [Chai, 2002]. Järjestelmän tehokas toiminta edellyttää myös, että sen on pystyttävä tunnistamaan kaikki objektit, joihin käyttäjä syötteillään viittaa. Nämä viittaussuhteet saattavat olla monimutkaisia ja ristiriitaisia, jolloin tulkinnan tekeminen on vaikeaa tai jopa mahdotonta. Milota [2004] esittää multimodaalisen komennon tulkitsemisen vaikeuden eräänlaisena etsintäongelmana, jossa järjestelmän on etsittävä ja valittava kaikkien mahdollisten syötetulkintojen joukosta oikea tulkinta.

Tunnistettaviin syötteisiin (katso kohta 3.1) keskeisesti liittyvä ongelma on tunnistamisvirheiden suuri määrä. Virheet vähentävät huomattavasti luonnollisten modaliteettien, kuten puheen, tehokkuutta ja vaativat käyttöliittymään jonkinlaisen virheiden korjausmenetelmän. Tunnistamiseen pohjautuvissa järjestelmissä käyttäjätyytyväisyys on kiinni sekä tunnistamisen tarkkuudesta että korjausdialogien monimutkaisuudesta ja ylipäänsä virheiden korjauksen tehokkuudesta [Ao *et al.*, 2007]. Puhesyötteiden kohdalla tunnistaminen perustuu sanastoihin, joihin sisällytetään käyttäjän tarvitsemat sanat ja niitä vastaavat ilmaisut. Tunnistin vertaa sarjaa todennäköisiä foneemeja sanoihin ja malleihin sanastossa ja palauttaa parhaan arvauksensa tai joissakin tapauksissa useita vaihtoehtoisia arvauksia.

Oviatt [2000] huomauttaa puheteknologian toimivan kohtuullisesti äidinkieltään puhuvien käyttäjien kohdalla, tekstiä luettaessa ja ideaalisissa laboratorio-olosuhteissa. Ongelmia aiheuttaa erityisesti puhe luonnollisessa ja spontaanissa vuorovaikutuksessa, käyttäjällä oleva aksentti sekä aito ympäristö, jossa tavallisesti esiintyy vaihtelevia melutasoja (aiheuttavat noin 20-50 %:n pudotuksen tunnistamistarkkuudessa). Muuttumattomat äänilähteet voidaan usein mallintaa ja prosessoida onnistuneesti (esimerkiksi liikenteen melu), mutta reaalimaailmassa on yhä paljon ääniä, joiden ilmestymistä ei voida ennakoida. Syötteen tulkintaa vaikeuttaakin näin sekä itse melu että käyttäjien tapa puhua eri tavalla meluisissa olosuhteissa. Käyttäjät myös puhuvat itsekseen ja käyttävät paljon epäolennaista puhetta, jolloin järjestelmän on pystyttävä erottamaan puhuuko käyttäjä itsekseen, muille henkilöille vai järjestelmälle [Oviatt and Lunsford, 2005].

Myös katse- ja elesyötteet ovat ongelmallisia tunnistamisen kannalta. Zhang *et al*. [2004] luettelevat viisi katsesyötettä koskevaa ongelmaa: tietyillä näytön alueilla silmänliikettä ei voida tarkasti seurata, vaikeus kalibroida katseenseurantalaite tietyille käyttäjille, kalibraation heikkeneminen ajan myötä, katseenseurantalaitteen heikko resoluutio ja päänliikkeiden sekä vartalon liikkeiden tiukka rajoittaminen. Näiden ongelmien takia silmänliikettä ei välttämättä pystytä tarkasti seuraamaan ja käyttäjän katseen saatetaan olettaa olevan muualla kuin missä se itse asiassa onkaan, varsinkin jos näytöllä olevat objektit on sijoiteltu tiheästi. Yhdistettäessä puhesyöte visuaalisiin kuvauksiin on hyvä huomata, että adjektiivien (suuri, pieni) tai tilasuhteiden (vasemmalla, oikealla) visuaalinen merkitys on luonnostaan epäselvä, sillä kuvaus on aina riippuvainen kontekstista, käyttäjien oppimista tavoista ja mielentilasta [Käster *et al*., 2003].

Holzapfel *et al*. [2004] huomauttavat eleiden voivan olla ristiriitaisia siinäkin tapauksessa, että tunnistaminen itsessään onnistuisi täydellisesti. Esimerkkinä he käyttävät ikkunaa ja lamppua ikkunan edessä, jolloin osoitusele lamppuun voi viitata joko lamppuun itseensä tai ikkunaan. Tämän ristiriidan ratkaiseminen vaatii toiselta modaliteetilta saatavaa vahvistavaa informaatiota. Tutkimuksessaan Holzapfel *et al*. havaitsivat, että 52 %:a syötteiden (ele ja puhe) yhdistämisen virheistä johtuivat eleen tunnistamisen täydellisestä epäonnistumisesta ja 22 %:a siitä, että eleen tarkoitusta ei voitu tunnistaa riittävän tarkasti. Aikarajoitteiden takia eleitä ja puhetta ei voitu yhdistää 9 %:ssa tapauksista ja näissä epäonnistuminen johtui suurimmaksi osaksi eleiden havaitsemisen vääristä aikaleimoista. Loput virheet selittyivät puheentunnistamisen ongelmilla.

Viittaussuhteet multimodaalisissa järjestelmissä voivat vaihdella yksinkertaisista monimutkaisiin ja tarkoista ristiriitaisiin. Käytännössä viittaussuhteiden ratkaiseminen (*reference resolution*) tarkoittaa vastineiden etsimistä esimerkiksi pronomineille: jos käyttäjän puhesyöte on muotoa ”siirrä tämä”, niin tehtäväksi tulee selvittää mihin pronomini ”tämä” viittaa. Chai *et al*. [2004] toteavat viitteiden selvittämisen vaativan informaation yhdistämistä sekä multimodaalisista syötteistä että vuorovaikutuksen kontekstista, johon kuuluu muun muassa vuorovaikutuksen historia ja sovellusalueen vaikutus. Hankalaa viittaussuhteiden ratkaisemisesta tulee, kun syötteet sisältävät useita viiteilmaisuja (esimerkiksi yhteen sanalliseen viiteilmaisuun liittyy useita erilaisia eleitä). Sekä syötteiden tunnistaminen että viittaussuhteiden oikea tulkinta on tärkeää multimodaalisen vuorovaikutuksen onnistumiselle.

4.5. Syötevirheiden käsittely

Multimodaalisiin syötteisiin liittyvät ongelmat edellyttävät järjestelmältä sopivaa menetelmää virhetilanteiden käsittelyyn. Odotusongelman yhteydessä (katso alakohta 4.4.1) käytiin jo läpi sen ratkaisemiseen tähtääviä toimenpiteitä, joten seuraavaksi keskitytään lähinnä muihin syötteiden käsittelyssä käytettäviin ratkaisumenetelmiin. Oviatt [2000] erottelee sekä käyttäjäkeskeisen että arkkitehtuuriin pohjautuvan virheiden käsittelyn. Käyttäjäkeskeinen virheiden käsittely liittyy modaliteettien käyttöön: käyttäjät välttävät

modaliteettia, jonka olettavat olevan tietyssä tilanteessa virheherkkä, käyttävät yksinkertaisempaa kieltä ja virheen sattuessa vaihtavat modaliteettia virheen tehokkaan ratkaisemisen varmistamiseksi. Arkkitehtuuriin pohjautuva virheiden käsittely puolestaan toteutuu hyvin suunnitelluissa ja optimoiduissa järjestelmissä, joissa kaksi syötesignaalia voivat selittää toinen toisensa.

Mankoff ja Abowd [Mankoff and Abowd, 1999] luokittelevat tunnistamiseen pohjautuvien käyttöliittymien virheiden käsittelyn viiteen päätutkimusalueeseen:

- Virheiden vähentäminen. Täysi virheettömyys on tuskin koskaan saavutettavissa tunnistettavien syötteiden kohdalla, mutta tunnistamisteknologiaa parantamalla voidaan pyrkiä vähentämään virheiden lukumäärää.
- Virheiden tunnistaminen. Ennen kuin järjestelmä tai käyttäjä voi toimia virheen korjaamiseksi, on jommankumman havaittava tapahtunut virhetilanne. Käyttäjä voi ilmoittaa järjestelmälle virheestä syötteellä ja järjestelmä auttaa käyttäjää löytämään virheet ilmoittamalla niistä tulosteessa. Automaattinen virheiden havaitseminen on mahdollista kynnysarvojen, sääntöjen ja tilastojen avulla.
- Virheiden korjaustekniikat. Virheiden käsittelytekniikat voidaan jakaa kolmeen kategoriaan, joista jokainen tarjoaa erilaisia keinoja virheiden korjaamiseen.
- Korjaustekniikoiden arvioiminen. Eri tekniikoita arvioimalla ja vertailemalla pyritään selvittämään niiden tehokkuus virheiden käsittelyssä.
- Kehitysvälineiden (*toolkit*) tuki. Kehitysvälineiden tavoitteena on tarjota uudelleenkäytettäviä komponentteja, jotka soveltuvat tavallisten ja samantyyppisten ongelmien ratkaisuksi. Käyttöliittymien virheiden käsittelyyn parhaiten sopiva kehitysväline olisi käytettävissä jokaisen virheherkän tilanteen aikana, jolloin näiden tilanteiden hoitaminen olisi suunnittelijan kannalta helpompaa.

Virheiden käsittelytekniikat Mankoff ja Abowd [Mankoff and Abowd, 1999] jakoivat kolmeen kategoriaan: oletusarvon valinta, selkeämpien syötteiden suosiminen ja ihmisten luonnollisten korjausstrategioiden jäljittely. Oletusarvon valintaan liittyy ”oikean” vastauksen valitseminen järjestelmän antamista vaihtoehdoista. Vastaavasti tehtävästä riippuen oletusarvo voidaan myös jättää kokonaan valitsematta tai esittää useita vaihtoehtoja, joista käyttäjän tulee itse valita oikea. Selkeämmät syötteet liittyvät edellä mainittuun käyttäjäkeskeiseen virheiden käsittelyyn: järjestelmä voi esimerkiksi virheherkän modaliteetin lisäksi tarjota luotettavamman syötekanavan. Ihmisten luonnollisia virheiden korjausstrategioita ovat muun muassa tauon pitäminen puheessa tai väärän ilmaisun toistaminen korjattuna ja kirjoitettaessa kirjaimen tai sanan poistaminen tai lisääminen aiemmasta tekstistä. Näitä korjauksia ei aina edes tietoisesti havaita niiden luontevuuden takia, joten niiden mallintaminen ja käyttö multimodaalisessa järjestelmässä voisi tehdä vuorovaikutuksesta aiempaan verrattuna huomattavasti sujuvampaa.

Multimodaalisissa järjestelmissä toisiaan täydentävät modaliteetit tulisi yhdistää siten, että ne tukevat toisiaan virheiden ratkaisussa (*mutual disambiguation*) ja parantavat näin järjestelmän toimintaa [Oviatt, 1999b]. Esimerkkinä tällaisesta voisi olla tilanne, jossa käyttäjä antaa puhekomennon ”talot”, jonka puheentunnistus virheellisesti tunnistaakin yksikkömuotoiseksi komennoksi ”talo”. Jos käyttäjä on kuitenkin valinnut osoituseleen avulla näytöltä useampia talo-objekteja, voi modaliteettien yhdistäminen johtaa oikeaan monikkomuotoiseen tulkintaan. Oviatt [1999b] analysoi multimodaalista järjestelmää, jossa yksi kahdeksasta järjestelmän tunnistamasta komennosta tulkittiin oikein vasta modaliteettien keskinäisen kompensaation jälkeen, vaikka tunnistimet olivat epäonnistuneet käyttäjän syötetavoitteen tunnistamisessa. Huolimatta siitä, että syötteen oikea tulkinta on tunnistimen n-best -listalla alempana kuin väärä tulkinta, voidaan oikeaan tulkintaan loppujen lopuksi päätyä, jos se on vaihtoehdoista ainut, joka on yhdistettävissä toisen modaliteetin syötteen kanssa.

Yksi keino ohjata käyttäjän syötettä vuorovaikutuksen aikana on ilmiö nimeltä kielellinen mukautuminen (*linguistic convergence*): ihmisten puheen ja kielimallien taipumus muuttua samanlaisiksi kuin kommunikaation toisella osapuolella. Tätä hyödyntämällä käyttöliittymä voi huomaamattomasti ohjata käyttäjän syötettä mukautumaan järjestelmän tulosteiden piirteisiin, jolloin voidaan välttää selkeiden rajoitteiden (ohjeet, harjoitus, virheilmoitukset) määrääminen käyttäjän toiminnalle [Oviatt and Lunsford, 2005]. Chai *et al*. [2004] käyttivät viittaussuhteiden selvittämiseen todennäköisyyksiin perustuvaa menetelmää, joka yhdisti ajalliset, semanttiset ja kontekstiin perustuvat rajoitteet ja johti niistä todennäköisimmän tulkinnan, joka parhaiten täytti kaikki rajoitteet. Näiden lisäksi virheitä käsiteltäessä voidaan hyödyntää esimerkiksi tilastomenetelmiä, erilaisia painotuksia modaliteettien kesken tai yksinkertaisesti kysyä käyttäjältä selvennystä ristiriitaiseen syötteeseen.

5. Fuusiomenetelmiä

Keskeisessä osassa multimodaalista järjestelmää on fuusioprosessi, joka yhdistää eri modaliteeteilta saadut syötteet yhdeksi merkitykselliseksi yhteistulkinnaksi. Fuusioprosessin toteutus ei ole aivan yksinkertainen tehtävä, vaan huomioon on otettava aiemmin mainitut käyttäjään, modaliteetteihin ja syötteisiin liittyvät ominaisuudet ja ongelmat. Lisäksi oma vaikutuksensa on sillä, mikä on järjestelmän käyttötarkoitus ja millaisia suoritettavat tehtävät ovat. Multimodaalisten syötteiden fuusiota varten onkin kehitelty useita erilaisia menetelmiä, jotka vaihtelevat sekä fuusiotekniikan (esimerkiksi sääntöihin pohjautuvat ja tilastolliset menetelmät) että tason (piirre- vai käsitteellinen taso) mukaan.

Seuraavaksi käydään läpi muutamia fuusiomenetelmiä. Tarkastelu aloitetaan unifikaatioon pohjautuvilla ratkaisuilla, joista siirrytään kontekstia ja semanttisia verkostoja hyödyntäviin fuusiomenetelmiin. Lopuksi käsitellään aiempia fuusiotekniikoita yhdistäviä hybridimenetelmiä sekä joitakin erityisratkaisuja. Menetelmien yleisen kuvauksen ohella huomiota kiinnitetään myös siihen miten ne yrittävät ratkaista syötteiden käsittelyn ongelmia (odotusongelma ja epäselvät syötteet). Ellei toisin ole mainittu, kaikissa esiteltävissä fuusiomenetelmissä syötteiden yhdistäminen tehdään käsitteellisellä tasolla, jolloin syötteitä on jo voitu esiprosessoida signaali- ja piirretasolla.

5.1. Unifikaatio

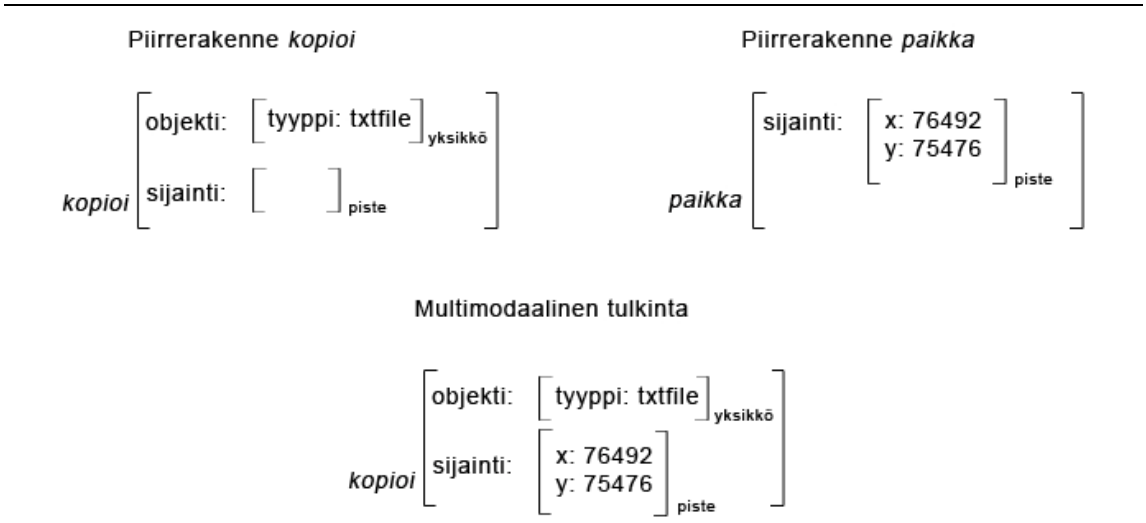
Tietojenkäsittelytieteessä unifikaatio ymmärretään operaatioksi, joka määrittää kahden osittaisen tietoyksikön yhtäpitävyyden. Jos tietoyksiköt ovat toisiaan vastaavia, on tuloksena niiden yhdistäminen yhdeksi yksiselitteiseksi tietoyksiköksi. Osittaisten tietojen keskinäinen vertailu tapahtuu erikseen määriteltyjen erityisten sääntöjen ja rajoitusten avulla. Multimodaalisten järjestelmien yhteyteen unifikaatio fuusiomenetelmänä sopii, sillä niissä syötteet ovat tavallisesti saman kokonaisuuden palasia, jotka tulee yhdistää yhdeksi tietokoneen käsiteltäväksi syötteeksi.

Johnston *et al*. [1997] toteuttavat multimodaalisten syötteiden fuusion tyypitettyjen piirrerakenteiden (*typed feature structure*) unifikaatiota käyttäen. Tyypitetty piirrerakenne kuvaa käsitteen, joka ilmaisee edustamansa kokonaisuuden ja joukon muita erottavia piirteitä. Rakenne on rekursiivinen, sillä kunkin erottavan piirteen arvona voi olla muuttujien sijasta toinen piirrerakenne. Piirrerakenteet esittävät modaliteettien syötteet semanttisina rakenteina, joka osaltaan helpottaa osittaisten tietoyksiköiden määrittelyä: syöte kuvataan vajaana piirrerakenteena, jonka tiettyjä piirteitä ei ole määritelty. Lisäksi rajoitteiden asettaminen fuusiolle helpottuu, kun piirteen arvon tyyppi voidaan ennalta määrätä.

Kun jokaiselle syötteelle on annettu oma tyypitetty piirrerakenteensa, ne siirretään integraatioyksikölle (usein erikoistunut agentti), joka etsii erillisten syötteiden parhaan

mahdollisen tulkinnan. Syötteiden uni- tai multimodaalisuuden selvittämiseksi eri modaliteettien syötteet merkitään joko täydelliseksi tai osittaisiksi riippuen siitä, tarvitaanko yhdistämistä. Jos syöte tarjoaa kokonaisen komentomäärittelyn, se merkitään täydelliseksi eikä se näin ollen vaadi yhdistämistä toisen syötteen kanssa. Jos taasen yksittäinen syöte täytyy yhdistää toiseen syötteeseen suoritettavan komennon aikaansaamiseksi, merkitään syöte osittaiseksi. Odotusongelman ratkaisemiseksi integraatioyksikkö käyttää aikaikkunaa, joka määrittää sen, ovatko syötteet ajallisesti yhdistettävissä. Aikaikkunan pituus mukailee modaliteettien synkronisaatiosta tehtyjä tutkimuksia. Syötteiden ristiriitaisuudet puolestaan yritetään välttää piirrerakenteiden tyyppirajoitusten ja modaliteettien keskinäisen kompensaation avulla.

Esimerkkinä voidaan ajatella tehtävää, jossa käyttäjän on kopioitava tietty objekti näytöltä. Ensin käyttäjä antaa suullisen käskyn ”kopioi” ja tämän jälkeen osoittaa kohdetta, joka tulisi kopioida. Syötteille annetut piirrerakenteet voisivat olla *kopioi* puheilmaukselle ”kopioi” ja *paikka* osoitetun kohteen sijainnille näytöllä. Koska sekä puhe että ele molemmat antavat vain osittaisen tulkinnan, on puheen piirrerakenne yhdistettävä unifikaation kautta tyypiltään *pisteeksi* merkittyihin sijaintiarvoihin (sovelluksessa saattaisi olla myös piirrerakenteita, joiden tyyppi olisi *viiva* tai *ympyrä*). Koska puhesyötteen tulkinta vaatii piirteeksi tyypiltään muotoa *piste* olevan sijainnin, ei unifikaatio elesyötteen muun tyyppisten piirrerakenteiden kanssa onnistu (Kuva 7).



Kuva 7. Kaksi erillistä piirrerakennetta yhdistetään multimodaaliseksi tulkinnaksi.

Tyypitettyjen piirrerakenteiden unifikaation rajoitteiden takia Johnston [1998] myöhemmin täydensi fuusiomenetelmäänsä tukemaan useiden syötteiden yhdistämistä: aiempi ratkaisu mahdollisti vain yksittäisen syötteen yhdistämisen toiseen. Useita ulottuvuuksia käsittävää taulukkojäsennintä hyödyntämällä unifikaatio voi tukea multimodaalisen vuorovaikutuksen spatiaalisia, ajallisia ja akustisia ulottuvuuksia mahdollistamalla niiden kesken jakautuneiden elementtien yhdistämisen. Multimodaaliset yhdistämisstrategiat määritellään unifikaatioon pohjautuvassa kielioppiformalismissa, jota taulukko-

jäsennin tulkitsee. Yhdistettävä informaatio ja rajoitteet voidaan ilmaista tekniikoilla, jotka sallivat yksittäisten syötteiden kuvaamisen tavalla, joka niiden luonteeseen parhaiten sopii.

Bin *et al*. [2000] toteuttivat unifikaatioon pohjautuvan syötteiden fuusion yhdistetyn piirrejoukon (*complex feature set*) avulla, joka on luonnollisen kielen ymmärtämiseen tähtäävä metodi. Lähtöajatuksena heillä olikin, että fuusion suorittaminen onnistuneesti vaatii esimerkiksi puhesyötteen kohdalla sen merkityksen ymmärtämistä eikä pelkkä merkkijonoiksi muuttaminen riitä. Yhdistetty piirrejoukko antaa jokaisen modaliteetin osittaiselle syötteelle yhtenäisen esitystavan (kaikille käytetään samoja muuttujia, kuten modaliteetti, aika ja tyyppi), jotka voidaan unifikaation kautta yhdistää. Perusoperaation sijasta Bin *et al*. käyttävät laajennettua unifikaatiota, jossa syötteiden ei tarvitse olla täysin yhtäläisiä, vaan niille riittää tietynasteinen yhteensopivuus: esimerkiksi kahden osittaisen syötteen aikaleimat eivät välttämättä ole yhtenevät, mutta silti niin lähekkäin toisiaan, että syötteet tulisi yhdistää. Syötteiden ristiriitaisuudet ja puutteet menetelmä käsittelee käyttämällä syntaktisia ja semanttisia rajoitteita, jotka voidaan ilmaista piirteiden välisinä suhteina.

Myös Sun *et al*. [2006a, 2006b] käyttivät unifikaatiota fuusiomenetelmissään QuickFusion ja MUMIF (*Mountable Unification-based Multimodal Input Fusion*). Ne molemmat käyttävät tyypitettyjä piirrerakenteita esittämään modaliteettien syötteitä. QuickFusion hyödyntää aikaleimojen sijaan syntaktista informaatiota odotusongelman ratkaisemiseksi, jolloin järjestelmä pystyy reagoimaan nopeammin käyttäjän toimiin. Omaan kielioppiinsa perustuva fuusioyksikkö päättää, onko saatu komento täydellinen vai ei. Jos komento on puutteellinen, odotetaan uusia syötteitä ja heti, kun oikeat lisäkomennot on saatu, suoritetaan fuusioprosessi. Virheiden käsittelyssä QuickFusion hyödyntää modaliteettien keskinäistä kompensaatiota ja toisaalta myös rajoittaa komentojen vapautta kielioppinsa määrittelyjen kautta. MUMIFin tavoitteena on pyrkiä uudelleenkäytettävään ja joustavaan fuusioon. MUMIF-yksikkö ottaa vastaan puhe- ja elesyötteet, jäsentää ne erikseen ja tämän jälkeen jäsennysten tulokset yhdistetään käyttäen multimodaalisen kieliopin integraatiosääntöjä, jotka kertovat millaiset syötteet voidaan yhdistää keskenään. Multimodaalisen kieliopin syntaksi on erillinen fuusiologiikasta ja toteutustavasta, jolloin sovelluksen vaihtuessa vain kielioppi tarvitsee vaihtaa. Lisäksi monimutkainen kielioppi voidaan tarvittaessa jakaa yksinkertaisempiin osiin sovelluksen eri osa-alueilla.

Unifikaatioon pohjautuvat fuusiomenetelmät käyttävät odotusongelman ratkaisemiseen pääsääntöisesti aikaleimoja, joiden perusteella voidaan päätellä, esiintyvätkö syötteet ajallisesti riittävän lähellä toisiaan. Tämä tekniikka tuo mukanaan viiveen, jonka enimmäisaika usein määräytyy tutkimuksissa havaittujen modaliteettien synkronisaatioaikojen mukaan (noin 3-4 sekuntia). Unimodaalisen syötteen antanutta käyttäjää tämä saattaa turhauttaa, varsinkin jos syöte virheellisesti tulkitaan osittaiseksi, jolloin sovel-

lus jää odottamaan täydentävää syötettä. Integraatiosäännöistä tulisikin tehdä riittävän yksityiskohtaiset, jotta järjestelmä tietää, miten eri komentojen kohdalla tulisi toimia. Johnstonin [1998] muutoksia lukuun ottamatta menetelmien heikkoutena on myös niiden kyky käsitellä kerralla vain kahta yksittäistä syötettä (esimerkiksi puhe- ja elesyötettä), mikä tekee niistä sopimattomia sovelluksiin, joissa syötteitä annetaan yhdellä kertaa useita.

5.2. Kontekstin hyödyntäminen

Kontekstia hyödyntävät fuusiomenetelmät käyttävät syötteiden tulkitsemisen helpottamiseen vuorovaikutustilanteen käyttökontekstia ja tehtävähistoriaa. Näin tulkintojen tekemistä voidaan avustaa käyttämällä informaatiota käyttäjän aiemmista kommunikaatiovuoroista. Useista mahdollisista tulkinnoista paras valitaan tavallisesti käyttämällä multimodaalisiin sääntöihin perustuvaa viitekehystä ja luottamusarvoja, jotka määrittävät syötteiden keskinäiset suhteet.

Pfleger [2004] esittelee fuusiomenetelmän, joka pohjautuu ideaan siitä, että jokainen syötetapahtuma on tulkittava suhteutettuna sen lokaalisen vuoron kontekstiin (*local turn context*). Lokaalisen vuoron konteksti käsittää kaikki aiemmin tunnistetut unimodaaliset syötetapahtumat ja dialogitilan, jotka molemmat kuuluvat käyttäjän kommunikaatiovuoroon. Kun kommunikaatiovuoro on päättynyt, fuusiokomponentti tarvittaessa muodostaa yhdistetyn syötteen, joka todennäköisimmin ilmaisee käyttäjän tavoitteen. Tämän jälkeen aiemmat kontekstuaaliset edustumat poistetaan ja alustetaan uusi lokaalisen vuoron konteksti yhdessä alkavan kommunikaatiovuoron kanssa. Syötteen uni- tai multimodaalisuuden määräävät yhdessä sekä konteksti että erityinen kokoelma multimodaalisia integraatiosääntöjä.

Integraatiosääntöjen tulisi kietoutua yhteen toistensa kanssa, jotta ne parhaiten onnistuvat tulkitsemaan ja yhdistämään kontekstuaalista informaatiota. Pfleger jakaa säännöt kolmeen eri luokkaan:

- Modaliteettien synkronisaatiosäännöt, joiden tehtävänä on valvoa vuorovaikutuslaitteiden prosessointitilaa ja tilamuutosten tapahtuessa päivittää kontekstuaalisia edustumia.
- Multimodaalisten ilmausten tulkintasäännöt, jotka vaihtelevat sen mukaan, onko kyse aidosta multimodaalisesta tapahtumasta, toisiaan vahvistavista syötteistä vai ristiriitaisista tapahtumista. Aidon multimodaalisen tapahtuman yhteydessä sääntöjen tulee tunnistaa, kumpi modaliteeteista määrittää yhdistämisen toimintakehyksen ja kumpi on yhdistettävä syöte: esimerkiksi puhesyöte usein muodostaa sen taustan, johon elesyöte liitetään. Vahvistavien syötteiden kohdalla sääntöjen tulee tunnistaa ne, vaikka syötteet hieman eroaisivatkin toisistaan. Ristiriitaisissa tapahtumissa sääntöjen tulee valita erilaisista hypoteeseista se, joka parhaiten sopii sen hetkiseen dialogitilaan.

- Unimodaalisten ilmausten tulkintasäännöt, joiden avulla tunnistetaan unimodaaliset syötteet ja valitaan paras mahdollinen hypoteesi, joka vastaa dialogitilaa.

Parhaan tulkinnan löytämiseksi ja ristiriitaisuuksien ratkaisemiseksi fuusiomenetelmä käyttää luottamusarvoja järjestämään eri tulkintahypoteesit ja ilmaisemaan niiden sisältämien virheiden todennäköisyyden. Odotusongelma hoidetaan ajallisten rajoitusten avulla, jotka tarkkailevat eri syötteiden välisiä ajallisia suhteita. Pflegerin esittämä fuusiomenetelmä myös mukautuu helposti käyttäjän vuorovaikutustapaan, jos omaksuttu tapa pystytään tunnistamaan: vuorovaikutustapa voidaan sisällyttää kontekstiin ja näin parantaa myöhempää syötteiden tunnistamista.

Jos käyttäjä suorittaa aiemman esimerkin mukaista kopiointitehtävää (katso kohta 5.1), on tärkeää, että puhe- ja elesyöte annetaan samassa lokaalisen vuoron kontekstissa. Jos syötteiden väli on liian pitkä, voi jompikumpi tunnistimista päättää syötteiden kuuntelun ja raportoida time-out-tilanteen, jolloin sovellus prosessoinnin jälkeen alustaa uuden lokaalisen vuoron kontekstin. Vaikka dialogitila on tärkeässä roolissa syötteitä yhdistettäessä, saatetaan puhe ja kopioinnin kohteen osoittava ele tässä tapauksessa tulkita virheellisesti unimodaalisina syötteinä ja näin käsitellä väärien integraatiosääntöjen kautta.

Sekä Nigay ja Coutaz [Nigay and Coutaz, 1995] että Hurtig ja Jokinen [Hurtig and Jokinen, 2006] jakavat fuusioprosessin kolmeen osaan, jonka viimeisessä vaiheessa hyödynnetään kontekstia. Lisäksi Nigay ja Coutaz esittelevät erityisen tietorakenteen nimeltä sulatusuuni (*melting pot*), joille fuusio heidän menetelmässään suoritetaan. Fuusiomenetelmän ensimmäisessä vaiheessa toteutetaan mikrotemporaalinen fuusio, joka yhdistää syötteet, joiden aikavälit ovat joko rinnakkaisia tai päällekkäisiä. Toisessa vaiheessa suoritetaan makrotemporaalinen fuusio syötteille, jotka ovat vaiheittaisia tai joiden aikavälit eivät ole päällekkäisiä, vaikka ne kuuluvat samaan aikaikkunaan. Viimeisessä vaiheessa kontekstuaalinen fuusio yhdistää toisiinsa liittyvät syötteet käyttäen informaatiota sen hetkisestä kontekstista eikä aikaväleihin kiinnitetä enää huomiota: käyttäjä saattaa esimerkiksi antaa käskyn, poistua hetkeksi ja palata sitten jatkamaan kesken jäänyttä tehtävää. Uusi syöte yhdistetään joukkoon aiempia syötteitä, jos tuleva syöte täydentää jotakin aiemmista syötteistä. Odotusongelma ratkaistaan käyttämällä niin kutsuttua ahnetta strategiaa: syötteiden yhdistämistä yritetään jatkuvasti eikä tulevaa informaatiota varsinaisesti odoteta. Käyttäjä voi näin saada välitöntä palautetta, joskin ongelmana on väärien fuusioiden tapahtuminen.

Hurtig ja Jokinen [Hurtig and Jokinen, 2006] vuorostaan esittävät fuusiomenetelmän, jonka ensimmäisellä fuusiotasolla käytetään sääntöpohjaista algoritmia, joka etsii kaikki sallitut tavat yhdistää kahden syötteen sisältämä informaatio. Tämän seurauksena syntyy usein suuri joukko mahdollisia syötehypoteeseja, joita rajoittaa ainoastaan se, että syötetapahtumien ajallinen järjestys on säilytettävä. Toisella tasolla kaikille syötehypoteeseille annetaan tilastollisiin tietoihin perustuva painotus, jonka perusteella syöt-

teistä tehdään n-best -lista. Viimeisellä tasolla n-best -listan ensimmäistä syötehypoteesia verrataan dialogin sen hetkiseen tilaan ja apuna käytetään tietoja käyttäjän mahdollisista aikomuksista ja kontekstista. Jos syötehypoteesi ja dialogi eivät sovi yhteen, jatketaan n-best -listassa eteenpäin, kunnes sopiva syötehypoteesi löytyy.

Kontekstia hyödyntämällä voidaan unifikaatiota laajemmin tehdä päätelmiä syötteistä ja niiden mahdollisesta keskinäisestä yhteenkuuluvuudesta. Huomionarvoista on kyky helposti mukautua käyttäjän omaksumaan vuorovaikutustapaan, jonka huomioimisella voidaan saavuttaa suuria parannuksia multimodaalisten järjestelmien toimintaan. Vuorovaikutustavan pysyvyyden vuoksi kerran oikein tunnistettua tapaa voidaan jo melko alussa käyttää ohjaamaan sovelluksen toimintaa. Odotusongelman ratkaisu perustuu jälleen aikaleimoihin ja niiden rajoituksiin, poikkeuksena Nigayn ja Coutazin [Nigay and Coutaz, 1995] menetelmä, jonka ahne strategia tuo kuitenkin omat ongelmansa fuusioprosessiin. Toisaalta käyttäjän toimintaa seuraamalla voidaan arvioida syötteiden uni- tai multimodaalisuuden todennäköisyyttä tietyissä käyttökonteksteissa.

5.3. Semanttiset verkostot

Tässä luvussa käsiteltävien kahden kehysperustaisen (*frame-based*) semanttisen verkoston lisäksi mukana on myös yksi syötteiden jäsentimeen ja erityisiin fuusiosääntöihin perustava fuusiomenetelmä. Ellei menetelmien kohdalla ole toisin mainittu, odotusongelman ratkaisu ja virheiden käsittely pohjautuvat jo aiemmin esitettyihin ratkaisuihin. Näiden fuusiomenetelmien etuihin kuuluu niiden helppo laajennettavuus ja muokattavuus sekä uudelleenkäytön mahdollisuus.

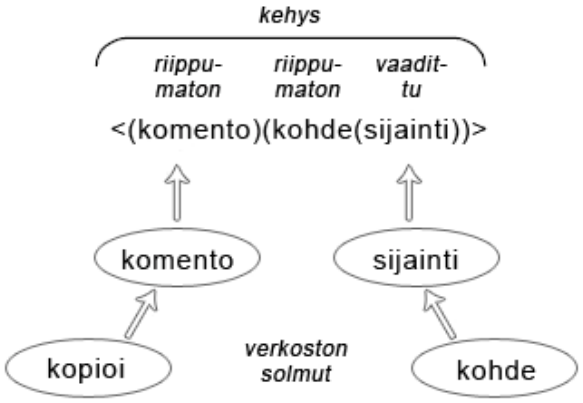
Tutkittuaan aiempia multimodaalisia fuusioratkaisuja Russ *et al*. [2005] sovelsivat niiden perusideoita omassa fuusiomenetelmässään. He käyttivät sanakirjatietorakenteeseen pohjautuvaa semanttista verkostoa, joka sisältää käytettävissä olevat syötteet, ympäristön ja tulkinnan. Toteutus on kehysperustainen ja sen vahvuus on käyttäjän syötteiden itsenäisyys kaikista ennalta määrätyistä vuorovaikutuksen muodoista: syötteitä rajoittaa vain verkoston käyttämä sanakirja, jota voidaan tarvittaessa laajentaa. Verkosto itsessään koostuu solmuista ja painotetuista yhteyksistä niiden välillä. Solmut sisältävät sekä termin (esimerkiksi sijainti, objekti, aika) että aktivaatioarvon.

Fuusioprosessi alkaa, kun syöte aktivoi semanttisen verkoston. Koska solmujen välillä on painotettu yhteys, yksittäisen solmun aktivaatio leviää verkostossa naapurisolmuihin painotuksista riippuen. Syötteiden mahdolliset tulkinnat määritellään kehyksinä, jotka koostuvat aukoista (*slots*), jotka taasen ovat yhteydessä verkoston solmuihin. Aukot jaetaan kolmeen osaan: ne voivat olla joko riippumattomia, vaadittuja tai attribuuttiaukkoja, joista jokainen sisältää erilaista informaatiota, joka täytetään, kun niihin yhdistetty solmu aktivoidaan (jos aktivaatioarvo on riittävän korkea). Jotta vaadituille aukoille voidaan löytää oikea syöte kaikista mahdollisista, on niiden tyypit määritettävä: riippumaton aukko vain ilmaisee, että toinen, tietyn tyyppinen syöte tarvitaan. Attribuutti-

aukot kertovat esimerkiksi objektin kohdalla siihen kuuluvista ominaisuuksista, joita voivat olla muun muassa tyyppi, sijainti ja koko.

Odotusongelman ratkaisemiseksi fuusiomenetelmä käyttää häivyttämistekniikkaa, joka hallitsee solmujen aktivaatiota: kun tietty aikajakso on kulunut, solmujen aktivaatioarvo laskee, kunnes se lopulta saavuttaa nollan eikä aktivaatiota enää ole. Tämän jälkeen saatuja uusia syötteitä ei enää käsitellä aiempiin liittyvinä. Häivyttämistekniikkaa käytettäessä kaikkia syötteitä ei ole pakko saada tiukan ennalta määrätyn ajan aikana ja häivyttämistä voidaan käyttäjistä sekä sovelluksesta riippuen myöhemmin nopeuttaa tai hidastaa. Virheiden käsittely hoidetaan modaliteettien omien n-best -listojen avulla, jotka järjestävät syötetulkinnat sen mukaan, miten todennäköisiä ne ovat annetussa tilanteessa. Lisäksi menetelmä käyttää kontekstista saatuja tietoja tulkintojen oikeellisuuden varmistamiseksi.

Palataan jälleen hetkeksi aiempaan esimerkkiin (katso kohta 5.1). Semanttisessa verkostossa puhesyöte ”kopioi” aktivoi solmun nimeltä *komento*, joka siirtää aktivaation eteenpäin kehyksen vastaavalle aukolle (komento), jolloin aukko täytetään saadulla syötteellä. Elesyöte aktivoi *sijainti*-solmun, jonka osalta kehys edellyttää vaaditun aukon (sijainti) täyttämistä tietyn tyyppisellä syötteellä (tässä tapauksessa tieto osoituseleen kohteesta). Semanttisen verkoston toimintaa havainnollistaa kuva 8.



Kuva 8. Semanttinen verkosto: solmut, kehys ja aukot.

Kehysperustaisen fuusiomenetelmän toteuttivat myös Flippo *et al*. [2003]. Ennen fuusioprosessia syötteet on sijoitettu aikaleimoilla varustettuun semanttiseen jäsennyspuuhun, joka on muutettava kehyksiksi dialoginhallintaa varten. Muutoksen suorittavat erityiset agentit, jotka ottavat osan jäsennyspuusta, suorittavat sille muutoksen ja täyttävät näin aukon semanttisessa kehyksessä (Kuva 9). Tarvittaessa agentit voivat hyödyntää kontekstista saatuja tietoja koskien esimerkiksi sovelluksen tilaa tai aiempaa dialogia. Jokainen agentti saattaa tuottaa useita mahdollisia ratkaisuja todennäköisyysarvoineen aukkojen täyttöä varten, joista fuusioyksikkö valitsee sopivan. Valintaa ohjaavat sekä ratkaisujen todennäköisyysarvot että agentille määritetty painotus tai luottamusar-

vo: jokaisen ratkaisun saamat pisteet summataan ja valituksi tulee se, jonka lopullinen pistearvon on korkein.

Kuva 9. Fuusioprosessin eteneminen jäsennyspuusta kehyksiksi [mukailtu Flippo *et al*., 2003].

Syötteiden ristiriitaisuuksien käsittelyn Flippo *et al*. jättävät sovelluksen toteuttajan vastuulle, joka viime kädessä vastaa siitä mikä modaliteetti ”voittaa” eli millaiset painotukset kukin modaliteetti saa. Toteuttaja voi painotusten avulla esimerkiksi määrittää elesyötteen voittamaan aina, kun objekti on hiiren osoittimen alla, mutta jos hiiren osoitin on tietyn välimatkan päässä objektista, voittaakin puhesyöte. Ratkaisemattoman ristiriidan yhteydessä käyttäjältä voidaan myös suoraan kysyä selvennystä. Odotusongelmaan ratkaisuna ovat jälleen aikakynnykset, varsinkin jos puhe ei toimi ohjaavana modaliteettina (tiedetään millaisia syötteitä odottaa).

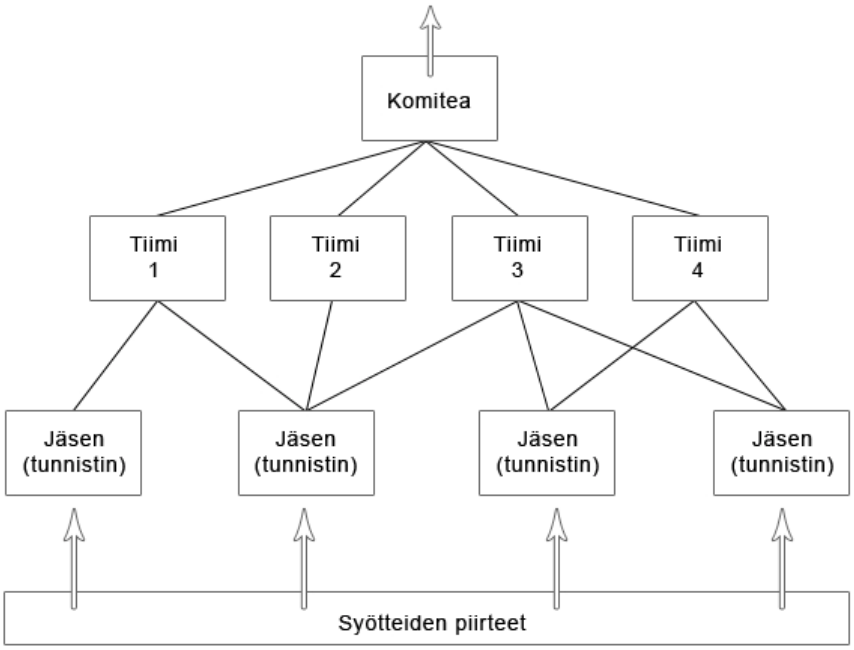
Semanttisten rakenteiden fuusio voidaan suorittaa myös ilman verkostoja. Holzapfel *et al*. [2004] perustavat semanttisten rakenteiden fuusion sovelluksesta riippumattoman jäsentimen ja sovelluksesta riippuvien fuusiosääntöjen varaan. Jäsennin käyttää apunaan syötemerkintöjä (*input tokens*) ja rajoitteita päättämään, mitkä syötteet voidaan yhdistää. Se käsittelee laajaa syötekokonaisuutta, johon voidaan lisätä uusia syötteitä ja vanhoja poistaa. Fuusio suoritetaan, kun syötemerkintöjen joukkoon voidaan sovittaa fuusiosääntö, joka koostuu rajoitteista ja johtaa lopulta muunnoskäsittelyyn, jonka tuloksena on yhdistetty syöte. Jokaisella fuusiosäännöllä on osa, joka määrittää, kuinka yhdistää useat syötemerkinnät ja osa, joka määrittää alkuarvot ja rajoitukset syötemerkinnöille. Virhetilanteissa hyödynnetään modaliteettien keskinäistä kompensaatiota ja virheellisesti tunnistetut syötemerkinnät, joihin fuusiosääntöjä ei voida lisätä, poistetaan syötekokonaisuudesta ennalta määrätyn ajan päätyttyä.

5.4. Hybridimenetelmät

Hybridimenetelmät hyödyntävät aiempia fuusiotekniikoita uusien fuusiomenetelmien kehittämisen pohjana. Kahden tai useamman vanhan tekniikan yhdistäminen voi auttaa ratkaisemaan yksittäisiä tekniikoita vaivaavat ongelmat ja tuoda uuden näkökulman fuusioprosessin hoitamiseen. Hybridimenetelmät ovat uusi suuntaus multimodaalisten järjestelmien kehittämisessä ja tehokkuutensa vuoksi varteenotettava vaihtoehto perin-

teiselle syötteiden yhdistämiselle. Ongelmia tuottaa kuitenkin niiden laitteistolta vaatima laskennallinen teho, joka estää hybridimenetelmien käytön tietyillä sovellusalueilla.

Wu *et al*. [2002] yhdistävät keskenään tilastolliset prosessointitekniikat ja unifikaatioon pohjautuvan, piirrerakenteita yhdistävän, symbolisen menetelmän. Fuusiomenetelmä käyttää assosiatiivista taulua (*associative map*) sulkemaan pois niiden puhe- ja elesyötteiden piirrerakenteet, joita ei voida yhdistää keskenään: mahdottomien unifikaatio-operaatioiden toteutuminen voidaan näin tehokkaasti estää. Syötteiden tunnistamisesta ja käsittelystä vastaa tekniikka nimeltä Jäsenet-Tiimit-Komitea (*Members-Teams-Committee*), joka on kolmikerroksinen hajoita-ja-hallitse-arkkitehtuuri varustettuna useilla jäsenillä ja tiimeillä sekä yhdellä komitealla (Kuva 10). Jäsenet ovat yksittäisiä tunnistimia, jotka tuottavat laajan skaalan tunnistustuloksia ilmaistuna ehdollisina posterioritodennäköisyyksinä. Ne voivat kuulua useampaan kuin yhteen tiimiin, joille raportoivat tuloksensa. Tiimit lisäävät tuloksiin erilaisia painotettuja parametreja, joita komitea sitten tarkastelee: sen tehtäviin kuuluu määrätä syötteen lopullinen tunnistamistulos.



Kuva 10. Jäsenistä, tiimeistä ja komiteasta koostuva arkkitehtuuri syötteiden tunnistamista, käsittelyä ja yhdistämistä varten [mukailtu Oviatt *et al.*, 2000].

Syötteiden tunnistamisen yhteydessä jäsenet analysoivat niiden itsensä kannalta merkityksellisen syötejoukon piirteitä tietyn modaliteetin osalta ja tämän jälkeen muodostavat n-best -listan tuloksista yhdessä arvioitujen posterioritodennäköisyyksien kanssa. Tiimit koordinoivat ja painottavat eri modaliteeteilta saadut syötteet: ne antavat posterioriarvion jokaiselle multimodaaliselle komennolle. Komitea analysoi posteriorien empiiristä jakautumaa ja järjestää lopullisen n-best -listan, joka lähetään piirrerakenteet yhdistävälle unifikaatio-operaatiolle. Menetelmän etuna on sen tarkka syötteiden tunnis-

taminen, jolla on suora vaikutus fuusioprosessin onnistumiseen. Lisäksi sen avulla voidaan tehokkaasti käsitellä vaikeita tiedon mallintamisongelmia. Perinteisiin menetelmiin verrattuna rakenteet saattavat vaatia enemmän laskentatehoa ja muistitilaa, joskin tämä on kiinni siitä, miten paljon toimijoita järjestelmään otetaan.

Toisen aiempiin fuusiomenetelmiin pohjautuvan hybridimenetelmän esittävät Portillo *et al*. [2006]. He yhdistävät keskenään unifikaatioon pohjautuvan fuusion, johon kuuluu multimodaalinen kielioppi sekä ajalliset rajoitteet (perustuu pitkälti Johnston *et al*. [1997] ja Johnston [1998] esittämään menetelmään) ja dialogitasolla tapahtuvan fuusion, jossa yhdistetään eri vuorovaikutuskanavilta saatu informaatio. Hybridimenetelmä yhdistää molempien fuusiomenetelmien hyvät puolet, mutta välttää samalla turhan monimutkaisuuden tuomisen järjestelmään. Se sisältää multimodaalisen kieliopin ja ajalliset sekä modaaliset rajoitteet kuten ensimmäisessä menetelmässä, mutta antaa lopullisen yhdistämisen dialoginhallinnalle, jolloin huomioon voidaan ottaa toisen menetelmän käyttämä syötteiden keskinäisen yhtäläisyyden ratkaisevan päätösprosessin tarjoama lisäinformaatio.

Kuvattu hybridimenetelmä mahdollistaa myös fuusioprosessin optimoimisen: niissä tapauksissa, joissa ajalliset ja modaaliset rajoitteet sisältävä jäsennys ei ole ristiriitainen, on vain yksi validi jäsennystulos, jolloin dialoginhallinnan ei enää tarvitse suorittaa siihen liittyvää ylimääräistä prosessointia. Ristiriitaisten syötteiden kohdalla multimodaalinen kielioppi sallii mahdollisten syötehypoteesien välittämisen dialoginhallinnalle, joka valitsee sopivimman vaihtoehdon. Uuden fuusiomenetelmän etuna on myös lisämodaliteettien tuoman monimutkaisuuden käsittelyn helpottuminen ja mahdollisuus suorittaa kerralla useampi tehtävä multimodaalisesti, joka ilman dialogitasolla tapahtuvaa fuusiota olisi hankalaa. Ongelmia tuottaa jälleen fuusiomenetelmän vaatima laskennallinen teho, joka estää sen soveltuvuuden esimerkiksi mobiilijärjestelmiin.

5.5. Muita fuusiomenetelmiä

Koska kaikkia multimodaalisia järjestelmiä varten luotuja fuusiomenetelmiä ei ole mahdollista käydä tässä läpi, otetaan jo esiteltyjen menetelmien lisäksi esille vielä pari erityisratkaisua. Niistä ensimmäinen perustuu äärellistiloihin, toinen spatiaalisen informaation erityispiirteisiin ja kolmas evoluutioteorian periaatteisiin.

Äärellistiloja multimodaalisessa fuusiossa hyödyntävät Johnston ja Bangalore [Johnston and Bangalore, 2000], joiden ratkaisu on samalla vaihtoehtoinen, mutta tehokkaampi lähestymistapa Johnstonin [1998] aiemmalle fuusiomenetelmälle. He käyttävät yhtä äärellistä tilakonetta suorittamaan syötteiden jäsennyksen, tulkinnan ja fuusion. Käyttämällä tiettyjä yksinkertaistavia oletuksia moniulotteinen jäsentäminen ja tulkinta multimodaalisten kielioppien avulla voidaan toteuttaa käyttämällä painotettua äärellistä automaattia, joka käyttää kolmea nauhaa edustamaan puhe- ja elesyötettä sekä niiden yhdistettyä tulkintaa. Yksinkertaistuksia käytetään ajallisten rajoitusten suhteen. Esimerkiksi useiden elesyötteiden yhteydessä rajoitusten pääasiallinen toiminta liittyy

eleiden järjestyksen hallitsemiseen: Johnstonin ja Bangaloren fuusiomenetelmässä multimodaaliset kieliopit koodaavat järjestyksen, mutta eivät määritä selkeitä aikarajoituksia.

Fuusio suoritetaan kirjoittamalla syötteiden tulkinta kolmannelle nauhalle, jonka ketjuttaminen tuottaa yhden semanttisen edustuman multimodaalisista syötteistä. Syötteiden ristiriitaisuudet ratkaistaan modaliteettien keskinäisen kompensaation ja todennäköisyyksien avulla: alussa pyritään tunnistamaan elesyötteet ennen puhesyötettä, jonka tulkinnan hoitavaa kielimallia (*language model*) voidaan tämän jälkeen muokata tunnistettuja eleitä hyödyntämällä. Jos elesyötteet ovat ristiriitaisia, ne esitetään taulukkona, joka sisältää kaikki mahdolliset syötehypoteesit ja tulkinnat. Puhesyötteen avulla taulukosta valitaan todennäköisin tulkinta eleille. Hybridimenetelmiin verrattuna äärellistiloja käyttävä fuusio ei vaadi suurta laskennallista tehoa, jolloin sitä voidaan hyödyntää useilla sovellusalueilla.

Spatiaalisissa kyselyissä käytetään tavallisesti useita puhe- ja piirtoeleitä, joiden käsittelyyn perinteiset fuusiomenetelmät eivät ole sopivia. Yhden spatiaalisen kyselyn aikana annetut useat, miltei samanaikaiset syötteet voivat perinteisillä menetelmillä johtaa väärien syötteiden fuusioon tai oikeiden syötteiden hylkäämiseen, jotka taasen aiheuttavat sen, että tietokannasta etsitään olemattomia objekteja tai hyödyllistä informaatiota menee hukkaan. Lee ja Yeo [Lee and Yeo, 2005] toteuttivatkin spatiaalisia kyselyjä varten oman fuusiomenetelmän (*Spatial Information Integration Technique*), joka tulkitsee syötteiden merkityksen vasta, kun koko tilanäkymä on kokonaan kuvattu. Syötteiden sisältämä informaatio kuvataan kokonaisuutena, jotta kaikki syötteen osat otettaisiin huomioon. Sekä puhe- että piirtosyötteiden sisältämää spatiaalista informaatiota hyödynnetään syötteiden yhdistämisessä keskenään: koska objektien järjestystä tai aikaleimoja ei käytetä tunnistamaan objektien esiintymistä, on objektien sekoittaminen tilanäkymässä vaikeaa. Tällainen yhdistäminen tekee lopputuloksesta luotettavamman kuin muita parametreja käytettäessä.

Kun syötteet on kerätty, ne prosessoidaan käyttökelpoisen spatiaalisen ja ei-spatiaalisen informaation erottamiseksi. Tämän informaation perusteella rakennetaan kuvausrakenteet puhe- ja elesyötteille, jotka seuraavassa vaiheessa fuusioidaan yhdeksi kuvausrakenteeksi. Fuusioprosessin aluksi puhesyötteen spatiaalinen informaatio jaetaan topologiseksi ja suuntaukset käsittäväksi informaatioksi, jotka muutetaan vastaamaan piirtosyötteen spatiaalisten suhteiden merkintöjä. Eri syötteiden merkintöjä voidaan näin helposti vertailla keskenään ja tuottaa yhdistetty kuvausrakenne, johon lisätään ei-spatiaalinen informaatio puhesyötteestä.

Sen lisäksi, että spatiaalisten kyselyjen fuusiomenetelmä kykenee hoitamaan useiden syötteiden yhdistämisen, se sallii rajoittamattoman puhesyötteen ja vapaan piirtoliikkeen, jolloin kyseisiä modaliteetteja voidaan käyttää hyvin joustavasti. Käyttäjät eivät ole pakotettuja käymään läpi harjoittelua tai muistamaan ennalta määritettyjä toi-

mintoja. Näin fuusiomenetelmä sekä kannustaa että tukee luonnollista vuorovaikutusta ihmisen ja tietokoneen välillä. Se myös tukee syötteiden jälkikäsittelytekniikkaa, jossa syötteiden käsittely aloitetaan vasta kommunikaatiovuoron päätyttyä. Muut fuusiomenetelmät aloittavat syötteiden prosessoinnin jo käyttäjän kommunikaatiovuoron aikana, mikä saattaa helposti johtaa virhetilanteisiin (syöte esimerkiksi vahingossa jätetään käsittelemättä, kun se esiintyy väärässä paikassa kommunikaatiovuoron aikana). Jälkikäsittelytekniikka mahdollistaa syötteiden alkuperäisyyden ja järjestyksen säilymisen eikä niitä vahingossa tai tahallaan jätetä käsittelemättä.

Evoluutioteoriaan perustuvan multimodaalisen fuusiomenetelmän esittävät Althoff *et al*. [2002]. Tilastoihin pohjautuva menetelmä käsittää populaation keskenään kilpailevia yksilöitä, joista jokainen edustaa ratkaisua esitettyyn ongelmaan. Populaatio luodaan uuden syötteen saapuessa ja jokainen yksilö arvioidaan sen sopivuuden mukaan. Tämän jälkeen jotkut yksilöistä valitaan yhdistämistä varten ja yhdistämisen kautta syntyneille uusille yksilöille jälleen lasketaan sopivuus ja luottamusarvot. Koska koko populaation sopivuus on uusien yksilöiden myötä muuttunut, täytyy kaikkien yksilöiden sopivuus laskea uudelleen: yksilön sopivuus riippuu siis koko populaation sopivuudesta. Tätä jatketaan, kunnes populaatio on sulautunut yhteen (lähes kaikki yksilöt vastaavat toisiaan) ja järjestelmä voi tuottaa tarkoituksenmukaisen komennon. Muussa tapauksessa koko algoritmi toistetaan alusta. Jos kesken prosessoinnin mukaan tulee uusia syötteitä, voidaan yksilöihin lisätä uutta tietoa. Tämä mahdollistaa uusien tunnistamistulosten ja vuorovaikutuslaitteiden lisäämisen suoraan fuusioprosessiin. Vaikka geneettiset algoritmit eivät takaakaan parhaimman mahdollisen ratkaisun löytämistä, pystyvät ne löytämään sopivia ratkaisuja hyvin nopeasti ja soveltuvat useille sovellusalueille.

6. EMMA: merkintäkieli multimodaalisille järjestelmille

Multimodaalisten syötteiden merkitsemistä varten on olemassa useita erilaisia merkintätapoja, kuten aiemmin mainitut tyypitetyt piirrerakenteet ja semanttiset kehykset. Varsinaista yleisesti käytössä olevaa virallista standardia ei kuitenkaan ole, vaikka W3C (World Wide Web Consortium) onkin pyrkinyt standardoimaan multimodaalisten syötteiden esitystavan esittelemällä EMMAn. Yksityiskohtaisempaan tarkasteluun EMMA valittiin, koska käytännössä se on ainoa merkintäkieli, joka tukee myös syötteiden yhdistämisen määrittelyä. EMMAa ja muita multimodaalisten käyttöliittymien merkintäkieliä on tarkasteltu myös aiemmassa julkaisussa Malmberg [2006].

6.1. Mikä on EMMA?

W3C:n esittämä merkintäkieli multimodaalisille järjestelmille on EMMA (Extensible MultiModal Annotation markup language), joka on XML-pohjainen tiedonsiirtoformaatti vuorovaikutuslaitteiden ja sovelluksen vuorovaikutuksenhallinnan välille [Johnston *et al*., 2007]. EMMAn on kehittänyt W3C:n multimodaaliseen vuorovaikutukseen erikoistunut työryhmä, jonka päämääränä on luoda määrittelyjä, jotka mahdollistavat Internetin käytön multimodaalista vuorovaikutusta hyödyntämällä. Esitetty merkintäkieli onkin tarkoitettu käytettäväksi järjestelmissä, jotka tuottavat semanttisia tulkintoja eri modaliteeteilta tulleille syötteille, jotka voivat olla unimodaalisia, jaksoittaisia, samanaikaisia tai yhdistettyjä.

EMMAa voidaan käyttää merkitsemään puhetta, luonnollista kieltä, graafisen käyttöliittymän syötteitä ja digitaalista mustetta, joskaan varsinaisia rajoitteita modaliteettien suhteen ei ole. EMMA esittää syötteen tulkinnan joukkona elementtejä, joihin liitetään erilaisia annotaatiota ohjaamaan multimodaalisten syötteiden yhdistämistä. Annotaatiot määrittävät syötteille muun muassa luottamusarvot, aikaleimat ja syötekanavan (akustinen, visuaalinen vai taktinen). Lisäksi syöte voidaan merkitä tunnistamattomaksi (*uninterpretable*), jos sitä ei voida tunnistaa oikein tai jos syöte ei vastaa sille määritettyjä arvoja (esimerkiksi puhesyöte ei vastaa annettua kielioppia).

EMMAa tulisi käyttää standardina tiedonsiirtoformaattina multimodaalisen järjestelmän eri komponenttien välillä, joissa EMMA-dokumentti luodaan automaattisesti edustamaan käyttäjän syötettä. Syötteet käsitellään ja tulkitaan yksittäisinä eikä dialogin aikana kerättyinä syötekokoelmina. Merkintäkieli on suunniteltu riittävän yleisluontoiseksi, jotta se voisi Internetin lisäksi tukea mahdollisimman laajaa valikoimaa sovelluksia. Kaiken kaikkiaan EMMA tarjoaa joustavan ja sopivan rakenteen multimodaalisten syötteiden kuvaamisen, vaikka onkin vielä kehitysvaiheessa.

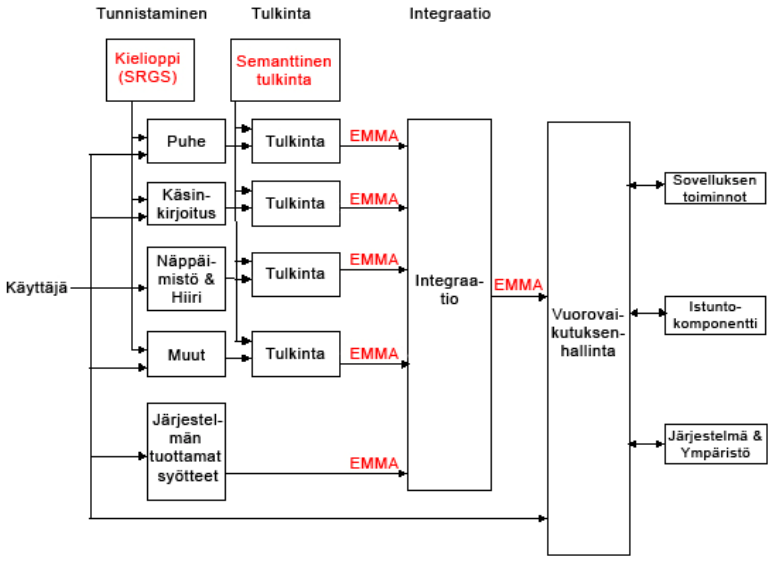
Tätä kirjoittaessa EMMA on W3C:n arvioitavana ja teknisen dokumentin asemassa, jolloin siihen voidaan yhä tehdä muutoksia – niin lisäyksiä kuin poistojakin. Uusin versio teknisestä dokumentista ilmestyi huhtikuussa 2007 ja esitteli useita muutoksia ja

selvennöksiä merkintäkielen rakenteeseen. Virallisen suosituksen saavuttaakseen EMMAn on kuitenkin vielä käytävä läpi useita tarkasteluvaiheita ja saattaa olla, ettei se koskaan saavuta standardin asemaa, jos W3C tai sovelluskehittäjät päättävät olla ottamatta sitä yleiseen käyttöön. Huolimatta keskeneräisyydestä on EMMAa jo käytetty syötteiden merkitsemiseen multimodaalisissa järjestelmissä, joista kerrotaan lisää kohdassa 6.3.

6.2. EMMA multimodaalisessa viitekehyksessä

W3C on esittänyt yksinkertaistetun viitekehyksen multimodaalisesta vuorovaikutuksesta, jota osaltaan voidaan hyödyntää hahmottaessa EMMAn toimintaa. Viitekehys pyrkii tunnistamaan multimodaalisten järjestelmien pääkomponentit ja ne merkintäkielet, jotka kuvaavat komponenttien tarvitseman informaation ja tiedonsiirron komponenttien välillä. Se kuvailee yleisesti käytössä olevat syöte- ja tulostemodaliteetit ja mahdollistaa uusien vuorovaikutuskanavien lisäämisen sitä mukaa kuin ne tulevat saataville. Pääpiirteissään viitekehyksen tärkeimmät toimijat ovat käyttäjä, syöte, vuorovaikutuksenhallinta (*interaction manager*), kontekstiin liittyvät komponentit ja tuloste. [Larson *et al.*, 2003]

Viitekehyksessä EMMA toimii syöteprosessorien ja vuorovaikutuksenhallinnan välillä: käyttäjän antama syöte voidaan muuttaa EMMA-dokumentiksi ennen kuin se toimitetaan vuorovaikutuksenhallinnalle (Kuva 11). Ennen tätä syötteitä voidaan käsitellä erillisessä integraatiokomponentissa, joka yhdistää eri modaliteettien syötteistä muodostetut EMMA-dokumentit keskenään ja lähettää yhteistulkinnan vuorovaikutuksenhallinnalle. Jos integraatiokomponenttia ei ole, suoritetaan dokumenttien yhdistäminen vuorovaikutuksenhallinnassa, jonka tehtävänä on ohjata kommunikaatiota käyttäjän ja järjestelmän välillä. Multimodaalisen viitekehyksen komponentteja ja EMMAn käyttöä siinä havainnollistaa kuva 11.



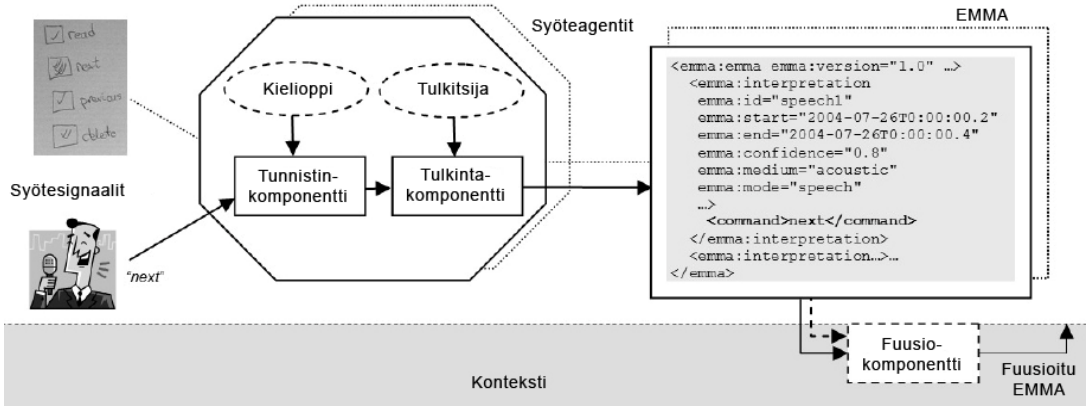
Kuva 11. Syötekomponentit ja EMMA [mukailtu Larson *et al.*, 2003].

Kuvassa 11 EMMAa tuottaviin komponentteihin kuuluvat puheentunnistimet, käsinkirjoitus, perinteiset syötetavat (hiiri ja näppäimistö) ja integraatiokomponentti. Muita mahdollisia ovat esimerkiksi luonnollisen kielen tunnistimet, katseenseurantalaitteet, DTMF-tunnistimet (*Dual-tone multi-frequency*) ja haptiset syötteet. EMMAa puolestaan käyttävät kuvan 11 mukaisesti vuorovaikutuksenhallinta ja integraatiokomponentti. Merkintäkieltä voidaan myös hyödyntää yleisenä semanttisena syötteen tulkintana, jota kuljetetaan mukana järjestelmässä ja täydennetään jokaisessa prosessointivaiheessa, vaikka tämä ei EMMAn varsinainen käyttötavoite olekaan. Lisäksi EMMAa voidaan tulevaisuudessa mahdollisesti käyttää välittämään abstraktia merkityssisältöä erityiselle käsittelykomponentille, joka kääntää sen luonnolliseksi kieleksi.

6.3. EMMA sovelluksissa

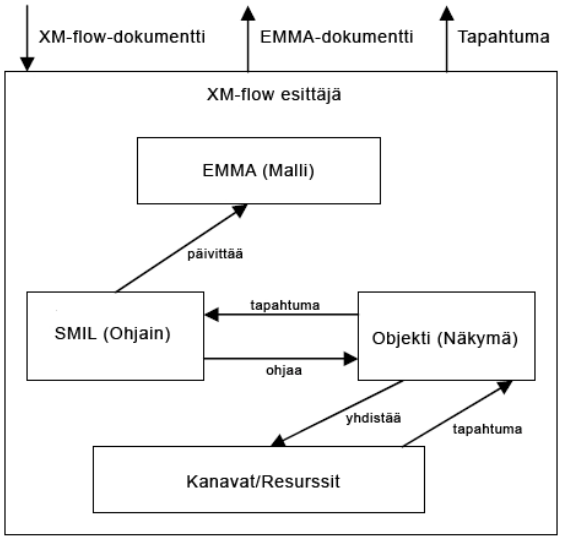
EMMA voi tulevaisuudessa mahdollistaa erilaisten vuorovaikutuslaitteiden käytön Internetin selailussa ja näin tehdä sen laajat resurssit uudenlaisten sovellusten saavutettaviksi. Tavoitteena on vuorovaikutus, joka olisi samanlaista huolimatta siitä, käyttääkö käyttäjä esimerkiksi puhelimen DTMF-äänimerkkejä, digitaalista mustetta vai ääniselainta. Tämä mahdollistaisi Internetin käytön yhä useammille ihmisille, kun selailu ei rajoitu pelkästään perinteisten modaliteettien käyttäjille. Toisaalta EMMAa voidaan käyttää laajemminkin multimodaalisten järjestelmien sovelluskehityksessä.

West *et al*. [2004] käyttivät EMMAa osana Nightingale-arkkitehtuuriaan, joka on kehitetty jokapaikan (*ubiquitous*) multimodaalisia sovelluksia varten. Arkkitehtuuri on agenttipohjainen ja siinä syöteagentit lähettävät syötteet EMMA-dokumenttien muodossa sovellusagenteille. Sovellus toimittaa syöteagenteille modaliteettikohtaiset kieliopit ja tunnistimet, jotka niitä hyödyntämällä tulkitsevat syötteet. Tämän jälkeen syötteet välitetään syöteagentissa olevalle semanttisen tulkinnan komponentille, joka muuttaa tunnistetut syötteet EMMA-dokumenteiksi, jotka tarvittaessa yhdistetään keskenään fuusiokomponentissa (Kuva 12).



Kuva 12. EMMA Nightingale-arkkitehtuurissa [mukailtu West *et al*., 2004].

Li *et al*. [2006] kuvaavat multimodaalisen arkkitehtuurinsa synkronisaatiomoduulin, joka vastaa syötteiden yhdistämisestä tulkitsemalla XM-flow-dokumenttia, joka sisältää EMMA ja SMIL (*Synchronized Multimedia Integration Language*) elementtejä. Arkkitehtuurissa dialoginhallinta tulkitsee XML-dokumenttia, joka määrittää kommunikaatiovuorot ja tämän jälkeen päättää mikä XM-flow-dokumentti tulee esittää vuorovaikutuksen tulkitsemiseksi. XM-flow-dokumentin esittämisen tuloksena on EMMA-dokumentti, joka yhdistetään dialoginhallintaan. Kuva 13 esittää XM-flow-arkkitehtuurin toiminnan ja EMMAn osuuden siinä. EMMA valittiin edustamaan multimodaalisia merkityksiä, jotta arkkitehtuuri olisi helposti siirrettävissä ja laajennettavissa.



Kuva 13. EMMA Malli-Näkymä-Ohjain-arkkitehtuuriin pohjautuvassa XM-flow-arkkitehtuurissa [mukailtu Li *et al*., 2006].

Laajennetun version EMMAsta tekivät Reithinger *et al*. [2005] SmartWeb-sovelluksessaan. SmartWeb antaa matkapuhelimen käyttäjille mahdollisuuden pyytää erilaisia palveluita, jotka on linkitetty esimerkiksi semanttiseen webbiin. Tulosteiden asianmukaisen esittämisen ja muualla järjestelmässä komponenttien väliseen viestintään käytetyn yleisen käyttöliittymäkielen takia Reithinger *et al*. päättivät laajentaa EMMAa ja lisäsivät siihen erityisiä rakenteita tulosteiden esittämistä varten. Nämä rakenteet käyttävät nimiavaruutta *swemma*. Tärkeimmät laajennukset ovat *result*-tunniste esittämään tunnistettua syötettä (ei varsinaista tulkintaa) ja *status*-tunniste SmartWebin sisäisen kommunikaatiotilan seuraamista ja eri prosessointitasojen tuloksia varten.

6.4. EMMA-dokumentin rakenne

EMMA-dokumentti voi sisältää kolmenlaista tietoa:

- Syötetieto (*instance data*) sisältää sovelluksesta riippuvan semanttisen tulkinnan käyttäjän tavoitteesta. EMMA tarjoaa yksinkertaisen rakennesyntaksin syötetietojen

ja tulkintojen kuvaamista varten. Syötteiden ristiriitaisuuksien takia EMMA-dokumentti voi sisältää useamman kuin yhden tulkinnan syötteestä. Tämä helpottaa syötteiden keskinäistä kompensaatiota sekä menneen tai nykyisen kontekstin hyödyntämistä tulkinnan teossa.

- Tietomalli (*data model*) määrittää syötteiden rakenteiden ja sisällön rajoitteet. Sen määrittely on valinnaista ja tavallisesti tietomalli on hyvin läheisesti kytköksissä sovellukseen.
- Metatieto (*metadata*) käsittää syötetiedon sisältämään informaatioon liittyvät annotaatiot. Annotaatioita käytetään kuvaamaan määriteltyä tietojoukkoa ja niihin kuuluvat esimerkiksi fuusioprosessin käyttämät aikaleimat, vuorovaikutuskanava ja tulkinnan luottamusarvot. EMMAn sisältämiä annotaatioita tulee pitää normatiivisina: jos ne ovat EMMAa käyttävän komponentin tuottamia, annotaatiot tulee esittää EMMAn syntaksia käyttäen.

Tietotyyppien esittely jättää tarkoituksellisesti syötteen tulkinnan mallintamisen epämääräiseksi ja antaa syötetietojen sekä tietomallien osalta vain suositteluja niiden esittämiseen. Tämä antaa sovelluskehittäjälle vapauden hyödyntää tulkinnassa haluamiaan tekniikoita eikä rajoita toteutusta tiettyihin malleihin: suosituksena on kuitenkin XML-pohjaisten merkintäkielten käyttäminen. Eniten huomiota kiinnitetään metatiedon määrittelemiin annotaatioihin, joilla onkin keskeinen osa syötteiden käsittelyssä.

6.4.1. Rakenne-elementit

EMMA-dokumentin juurielementti on *emma:emma*. Se esittää sekä käytössä olevan version EMMAsta että nimiavaruuksien määrittelyt. W3C:n esitys EMMAsta kuuluu kuvan 14 mukaiseen XML-nimiavaruuteen. EMMA-prosessorien täytyy kuitenkin tukea useita erilaisia XML-nimiavaruuksia ja lisäksi sovelluksesta riippuva merkintäkieli voidaan ilmaista joko avoimena tai määrittelemättömänä nimiavaruutena.

```
<emma version="1.0" xmlns="http://www.w3.org/2003/04/emma">
  ...
</emma>
```

Kuva 14. Nimiavaruuden määritteleminen.

Varsinainen syötteen tulkinta on elementissä *emma:interpretation*, jossa se esitetään sovelluksesta riippuvalla merkintäkielellä. Jos syötteen tulkinta epäonnistuu tai syötettä ei ole, jätetään elementti tyhjäksi. Ainoa elementin vaatima attribuutti on yksilöllinen id-tunniste, jonka avulla syötteen tulkinta voidaan EMMA-dokumentin sisällä erottaa muista tulkinnoista (Kuva 15).

```
<emma:interpretation id="001">
   ...
</emma:interpretation>
```

Kuva 15. Yksilöllinen id-tunniste erottaa tulkinnat toisistaan.

Säiliöt yhdelle tai useammalle elementille *emma:interpretation* ja muille säiliöelementeille ovat *emma:one-of*, *emma:group* ja *emma:sequence*. Ensimmäinen näistä on tarkoitettu kokoelmalle toisensa poissulkevia tulkintoja: esimerkiksi joukolle erilaisia tuloksia puheentunnistimelta. Toisensa poissulkevat syötteet johtuvat usein joko eri prosessointimenetelmistä tai ristiriitaisuuksista, kun sovelluksessa käytetään useita erilaisia tunnistimia samalle syötteelle tai kun sama syöte kerätään useammalla vuorovaikutuslaitteella (esimerkiksi monen mikrofonin käyttö). Käytettäessä elementtiä *emma:one-of* tulkinnat täytyy järjestää jonkin kriteerin mukaan paremmuusjärjestykseen: joko EMMAn elementtiä *emma:confidence* tai sovelluskohtaisia arvoja käyttämällä. Lisäksi elementtejä *emma:one-of* voidaan käyttää sisäkkäin, jolloin niiden yhteydessä täytyy ilmaista lähtökohta tulkintojen poissulkevuudelle attribuutin *disjunction-type* avulla.

Toinen säiliöelementti, *emma:group*, on tarkoitettu erillisille, mutta toisiinsa liittyville syötteille, jotka liitetään yhteen jonkin ryhmittelykriteerin perusteella. Elementin avulla voidaan ryhmitellä esimerkiksi samaan tavoitteeseen liittyvät puhe- ja elesyötteet. Tulkintojen ryhmittelykriteeri ilmaistaan käyttämällä elementtiä *emma:group-info*, joka voi viitata sovelluksessa itsessään tai sen ulkopuolella olevaan ryhmittelykriteereihin. Viimeistä säiliöelementtiä, *emma:sequence*, käytetään jaksoittaisten syötteiden tulkinnoille. Ajallisesti jaksoittaiset tulkinnat voidaan merkitä aikaleimojen avulla tarkan järjestyksen saavuttamiseksi.

EMMA mahdollistaa myös taulukoiden esittämisen elementillä *emma:lattice*. Taulukoiden avulla useat tunnistustulokset tai syötteiden tulkinnat voidaan esittää tiiviisti: ne kuvataan listana siirtymiä solmujen välillä. Jokaisella nimetyllä siirtymällä on alku- ja loppusolmut, joiden määritteleminen on EMMAn syntaksin mukaisesti pakollista. Itse elementti *emma:lattice* sisältääkin joukon elementtejä *emma:arc* ja *emma:node*, jotka koodaavat taulukkoesityksen käyttäjän syötteestä (Kuva 16). Esimerkiksi puhesyötteen kohdalla taulukkoesitys välttää tarpeen luetella kaikki mahdolliset sanajärjestykset ja mahdollistaa samalla annotaatioiden merkitsemisen syötteen yksittäisille sanoille.

```
<emma:lattice initial="1" final="3">
   <emma:arc from="1" to="2">…</emma:arc>
   <emma:arc from="2" to="3">…</emma:arc>
</emma:lattice>
```

Kuva 16. Taulukkoelementin alussa tulee määrittää ensimmäinen ja viimeinen solmu.

Merkkijonoliteraalit ilman sovelluksesta riippuvaa merkintäkieltä asetetaan elementtiin *emma:literal*. Jos esimerkiksi puhesyötteen semanttinen tulkinta olisi yksinkertaisesti ”kopioi” ilman ympäröiviä tunnisteita, tulee se sijoittaa mainittuun elementtiin.

6.4.2. Annotaatiot

EMMAn annotaatiot jaetaan annotaatioelementteihin ja annotaatioattribuutteihin. Annotaatioelementit sisältävät sisäisen rakenteen ja ne voivat esiintyä useammin kuin kerran toisten elementtien sisällä. Niillä määritellään muun muassa käytetty tietomalli (*emma:model*) ja kielioppi (*emma:grammar*). Yhteensä annotaatioelementtejä on seitsemän (tarkemmat kuvaukset EMMAn kaikista annotaatioista antavat Johnston *et al*. [2007]). Annotaatioattribuutit voivat esiintyä *emma:interpretation* elementissä ja osa niistä myös säiliö- ja taulukkoelementeissä sekä sovelluksesta riippuvan merkintäkielen elementeissä. Ne sisältävät annotaatiot muun muassa aikaleimoille, tunnistamattomalle syötteelle (*emma:uninterpreted*) ja käytetylle kielelle (*emma:lang*). Yhteensä erilaisia annotaatioattribuutteja on 34.

Multimodaalisen fuusioprosessin kannalta keskeisimpinä annotaatioattribuutteina voidaan pitää mediaan (*emma:medium*), kommunikaatiotapaan (*emma:mode*), toimintaan (*emma:function*), yhdistämiseen (*emma:hook*), luottamusarvoihin (*emma:confidence*) ja aikaleimoihin (esimerkiksi *emma:start*, *emma:end*) liittyviä attribuutteja. Media määritellään suljetuksi joukoksi arvoja, jotka määrittävät vuorovaikutuskanavan, joka voi olla akustinen, taktiilinen tai visuaalinen. Mediaa yksityiskohtaisempi on kommunikaatiotapa, joka ilmaisee tarkemmin tavan, jolla tietyn vuorovaikutuskanavan syöte annettiin: esimerkiksi taktiilista mediaa käytettäessä syöte voidaan antaa digitaalisella musteella tai hiiren osoituksilla. Samansuuntainen kommunikaatiotavan kanssa on toiminta, jossa syötteet luokitellaan suhteessa niiden kommunikatiiviseen tarkoitukseen. Esimerkiksi puhetta voidaan käyttää äänittämiseen, saneluun tai käyttäjän varmentamiseen. Näihin attribuutteihin liittyviä vaihtoehtoisia arvoja esitetään kuvassa 17.

```
emma:medium =[acoustic|tactile|visual]
emma:mode =[voice|dtmf|ink|gui|keys|video|photograph| ... ]
emma:function = [recording|transcription|dialog|verification| ... ]
```

Kuva 17. Annotaatioattribuuttien vaihtoehtoisia arvoja.

Luottamusarvoja käytetään merkitsemään syötteen laatua ja niitä voidaan antaa kaikenlaisille syötteille. EMMA-dokumentissa luottamusarvot ovat väliltä 0.0 ja 1.0, joista ensimmäinen edustaa pienintä mahdollista luottamusta. Luottamusarvoja ei tarvitse tulkita todennäköisyyksinä, vaan arvojen tulkinta on sovelluksesta riippuvaista. Aikaleimoihin kuuluvat annotaatiot alku- ja loppuajoille ja kestoille sekä aikavastineet (*time offsets*), jotka pohjautuvat sen hetkisen vuorovaikutuksen aikaleimoihin.

Syötteet, joiden sisältö tulee yhdistää toisen modaliteetin syötteen kanssa oikean tulkinnan saamiseksi, voidaan merkitä käyttäen attribuuttia *emma:hook*. Se kommunikaa-

tiotapa (*emma:mode*), jonka syöte tulee yhdistää parhaillaan käsiteltävään syötteeseen, on attribuutin *emma:hook* arvona. Näin ollen arvoksi sopii mikä tahansa aiemmin mainituista attribuutin *emma:mode* arvoista sekä lisäksi arvo *any*, jolloin yhdistettävä informaatio voi tulla mistä lähteestä tahansa. Huomattava on, että *emma:hook* ainoastaan merkitsee yhdistämistä vaativat syötteet, mutta ei siis suorita itse fuusioprosessia. Syötteiden fuusio voidaan suorittaa jollain yllämainituista fuusiomenetelmistä, kuten unifikaatiolla, joka sopivasti tarjoaa myös keinot syötteiden yhteensopivuuden tarkistamiseen. Seuraavassa kohdassa tarkastellaan lähemmin syötteiden merkitsemistä EMMAlla ja attribuutin *emma:hook* käyttöä.

6.5. EMMA käytännössä

Jotta EMMA-merkintäkielen käsittely ei jäisi tässä täysin teoreettiseksi, esitetään seuraavaksi yksinkertainen esimerkki EMMAn käytöstä tilanteessa, jossa kahdelta eri modaliteetilta saadut syötteet tulee yhdistää keskenään. Esimerkkitilanteena toimii jo aiemmin mainittu objektin kopiointi, jossa käyttäjä antaa sekä suullisen käskyn että osoittaa kopioitavaa kohdetta näytöllä. Jotta järjestelmä tietää mikä kaikista näytön objekteista sen on kopioitava, tulee kahdelta tunnistimelta saapuneet, erilliset EMMA-dokumentit fuusioida keskenään.

Kun käyttäjä antaa puhesyötteen ”kopioi”, siitä muodostaan EMMA-prosessorissa semanttinen kuvaus. Kuvaus käyttää attribuuttia *emma:hook* ilmaisemaan, että puhesyötteeseen tulee yhdistää elesyötteen (annetaan digitaalisena musteena) sisältö (Kuva 18). Selkeyden vuoksi seuraavista EMMA-dokumenteista on jätetty nimiavaruuksien määrittelyt pois ja sovelluksesta riippuvat tunnisteet ovat suomenkielisiä. Esimerkiksi tyyppi on sovellusriippuvainen tunniste, joka ilmaisee, että puhesyöte tulisi yhdistää osoituseleen eikä piirtoeleen kanssa.

```
<emma:emma>
  <emma:interpretation emma:medium="acoustic" emma:mode="voice">
    <komento>
      <toiminto>kopioi</toiminto>
      <objekti emma:hook="ink">
        <tyyppi>piste</tyyppi>
      </objekti>
    </komento>
  </emma:interpretation>
</emma:emma>
```

Kuva 18. Puhesyötteen EMMA-dokumentti.

Kun käyttäjä osoittaa kopioitavaa kohdetta, luodaan seuraava EMMA-dokumentti (Kuva 19). Attribuutti *emma:mode* kertoo käytetyn kommunikaatiotavan, joka tässä tapauksessa on digitaalinen muste.

```
<emma:emma>
  <emma:interpretation emma:medium="tactile" emma:mode="ink">
    <objekti>
      <kohde>file1.txt</kohde>
      <tyyppi>piste</tyyppi>
      <id>txt1</id>
    </objekti>
  </emma:interpretation>
</emma:emma>
```

Kuva 19. Elesyötteen EMMA-dokumentti.

Molemmat tarvittavat EMMA-dokumentit on nyt saatu ja ne voidaan yhdistää valitulla fuusiomenetelmällä. Jos käytetään unifikaatiota, se tunnistaa attribuutilla *emma:hook* merkityt elementit ja yrittää tämän jälkeen fuusioida ne vastaavalla kommunikaatiotavalla merkittyjen elementtien kanssa. Jos yhdistettävillä elementeillä ei ole ristiriitaisia arvoja alielementeissä tai attribuuteissa, unifikaatio-operaatio onnistuu. Unifikaation tuloksena syntyneestä EMMA-dokumentista poistetaan attribuutit *emma:hook* ja muutetaan attribuutin *emma:mode* arvoksi lista tavoista, joilla yksittäiset syötteet saatiin (Kuva 20).

```
<emma:emma>
  <emma:interpretation emma:medium="acoustic tactile" emma:mode= "voice ink">
    <komento>
      <toiminto>kopioi</toiminto>
      <objekti>
        <kohde>file1.txt</kohde>
        <tyyppi>piste</tyyppi>
        <id>txt1</id>
      </objekti>
    </komento>
  </emma:interpretation>
</emma:emma>
```

Kuva 20. Puhe- ja elesyötteen yhdistävä EMMA-dokumentti.

Annotaatioattribuutti *emma:hook* mahdollistaa sovelluksesta riippumattoman tavan ilmaista syöte, joka vaatii yhdistämistä toisen syötteen kanssa. Jos syötteiden fuusiossa käytetään unifikaation mukaista yleistä fuusiotekniikkaa, tulisi sovelluskehittäjien pystyä hyödyntämään samaa tekniikkaa useissa erilaisissa sovelluksissa ilman, että integraatiosääntöjä tai -logiikkaa tarvitsee muuttaa. Lisäksi attribuutin käyttö helpottaa multimodaalisten integraatiokomponenttien yhteistoimintaa, kun keskenään yhdistettävät syötteet on ilmaistu jo merkintäkielen tasolla.

7. Pohdintaa

Tämän tutkielman tavoitteena oli tarkastella modaliteettien käyttöön liittyviä erityispiirteitä ja esitellä multimodaalisten syötteiden yhdistämisessä käytettäviä fuusiomenetelmiä. Lisäksi mukaan otettiin EMMA, joka tarjoaa multimodaalisille järjestelmille oman (myöhemmin mahdollisesti standardoidun) merkintäkielen. Kaikki kolme osa-aluetta täydentävät toisiaan luontevasti eikä niiden tarkastelu yksittäisinä antaisi riittävän laajaa kuvaa multimodaalisten järjestelmien mahdollisuuksista ja haasteista. Tutkimusaiheena multimodaalisuus on sekä ajankohtainen että mielenkiintoinen.

Kiinnostus multimodaalisuutta kohtaan kasvaa jatkuvasti käyttäjien vaatiessa yhä luonnollisempia käyttöliittymiä, joiden käytön tulisi olla yhtä sujuvaa kuin minkä tahansa arkipäivän askareen suorittaminen. Tietokoneesta halutaan tehdä yhä ihmismäisempi, vaikka kyseenalaista onkin, miten pitkälle tällainen kehitys voidaan ja kannattaa viedä. Onko vuorovaikutusongelmien ratkaisuna keskusteluun ja päätöksentekoon kykenevä robotti vai tietokone, jonka kanssa kommunikointi ei poikkea käyttäjien jo oppimista vuorovaikutustavoista? Toisaalta tietokonejärjestelmien kehitystä tarkasteltaessa mielenkiintoista on se, miten kauan perinteiset käyttöliittymät ovat pitäneet pintansa. Siirtyminen niistä multimodaalisiin käyttöliittymiin vie aikansa eikä täyttä luonnollisuutta ihmisen ja tietokoneen väliseen vuorovaikutukseen saavuteta ehkä koskaan.

Yksi multimodaalisia järjestelmiä vaivaava ja niiden käytön luonnollisuutta häiritsevä ongelma on uusien modaliteettien syötteiden tunnistamisen vaikeus. Siinä, missä näppäimistön tai hiiren painallus on tavallisesti hyvin yksiselitteinen, voi puhesyötteeseen liittyä useita mahdollisia tulkintoja. Epämääräisyys johtuu pitkälti ihmisten erilaisista toimintamalleista ja siitä, ettei kaikkea käyttäytymistä pystytä hallitsemaan tai havaitsemaan tietoisesti. Täytesanat, yllättävät katseenliikkeet tai liiallinen voiman käyttö saattavat vääristää syötteen tulkintaa eikä käyttäjä välttämättä edes ymmärrä mistä virhe johtui. Tunnistamisongelmat eivät kuitenkaan saa johtaa uusien modaliteettien hylkäämiseen, vaan vuorovaikutusta tulee tehostaa uusia tunnistustekniikoita ja -laitteita kehittämällä. Ihmisten toiminnan mallintaminen pääpiirteissään ei enää riitä, sillä huomioon tulee ottaa kaikki kokonaisuuteen vaikuttavat osa-tekijät.

Tulevaisuudessa tietokoneiden tulisikin entistä paremmin mukautua käyttäjien toimintamalleihin. Vaikka käyttäjien välillä voikin olla huomattavaa vaihtelua, perustuu ihmisen toiminta lainalaisuuksiin, joita voidaan hyödyntää järjestelmiä suunniteltaessa. Käyttäytymistä ennakoimalla ja kontekstia tarkistelemalla voidaan sekä saavuttaa luotettavampia tunnistustuloksia että reagoida nopeasti annettuihin syötteisiin. Käyttäjänsä vuorovaikutustapoihin mukautuva järjestelmä lisäisi huomattavasti vuorovaikutuksen miellyttävyyttä ja tehokkuutta. Tämä merkitsisi samalla yhä luonnollisempaa vuorovaikutusta, kun edellytyksenä ei enää ole jokaisen käyttäjän sovittaminen samanlaiseen

muottiin. Vaikka onkin jo olemassa järjestelmiä, jotka käyttävät käyttäjän toimintamalleista kerättyä informaatiota hyväkseen, on kehitys tällä saralla vielä alkutekijöissään.

Toinen multimodaalisten järjestelmien ongelma on erilaisten syötteiden yhdistäminen keskenään. Yhtä ainoaa ratkaisua ei ongelmaan ole, vaan tässä tutkielmassa esitetyt fuusiomenetelmät edustavat useaa suuntausta multimodaalisten syötteiden yhdistämiseen. Yksittäisen menetelmän toimivuus (fuusioprosessin onnistuminen) on suurelta osin kiinni siitä, millaisella sovellus- ja tehtäväalueella sitä käytetään ja miten hyvin sovelluksen suunnittelija on osannut ottaa huomioon eri modaliteetteihin liittyvät erityispiirteet. Tulevien multimodaalisten järjestelmien kannalta eniten kiinnostusta herättävät erilaiset hybridimenetelmät, jotka yhdistelevät aiempia fuusiomenetelmiä. Ne tarjoavat vanhoja menetelmiä tehokkaammat fuusiotekniikat ja pystyvät eri tekniikoita yhdistelemällä selvittämään aiemmat ongelmakohdat. Ongelmallista useimpien fuusiomenetelmien kohdalla on kuitenkin niiden rajallinen kyky käsitellä ja yhdistää vain kahden, usein ennalta määrätyn, modaliteetin (tavallisesti puhe ja eleet) syötteet. Syynä tähän on osittain syötteiden merkintään käytetyt tavat, joissa ominaisuuksia on vain tiettyjä modaliteetteja varten.

Ratkaisuna multimodaalisten syötteiden merkintätapojen hajanaisuuteen on EMMA. Tällä hetkellä multimodaaliset järjestelmät saattavat olla sekamelska erilaisia merkintäkieliä: omansa jokaiselle modaliteetille ja syötteiden yhteistulkinnalle. Tämä vaikeuttaa sekä niiden laajennettavuutta että tekniikoiden yleiskäyttöisyyttä. Toisaalta EMMAkaan ei ennen yleistymistään juuri ole avuksi, sillä sovelluskehittäjät saattavat joko unohtaa koko merkintäkielen olemassaolon tai kokea sen käytön turhana. Myös EMMAn tämän hetkinen keskeneräisyys on sen käyttöä rajoittava tekijä. Uusimmassa teknisessä dokumentissaan W3C ilmoittaa useiden elementtien ja annotaatioiden olevan poistouhan alla – joukossa myös tässäkin tutkielmassa käsitelty attribuutti *emma:hook*. Lisäksi vanhoja määrittelyjä saatetaan yhä muuttaa ja uusia lisätä. Kaikista huolimatta EMMA on tarvittu lisä multimodaalisiin järjestelmiin, joiden laaja kirjo kaipaa yhteisiä merkintätapoja ja näin parempaa kokonaisuuden hallintaa.

Käytännössä multimodaaliset järjestelmät ovat vielä tutkimustyön alla laboratorioissa ja jotkut vuorovaikutuslaitteista ovat liian kalliita tavallisten käyttäjien hankittavaksi. Siksi aihepiiri saattaakin tuntua monesta hyvin kaukaiselta, vaikka todennäköistä on, että tulevaisuudessa yhä useammat käyttäjät hyödyntävät multimodaalisuutta ollessaan vuorovaikutuksessa tietokoneensa kanssa. Tämän hetken sovelluskehityksen tehtävänä on vielä etsiä ja kokeilla erilaisia malleja, joilla toteuttaa mahdollisimman luonnollinen vuorovaikutus ihmisen ja tietokoneen välille. Tässä työssä auttavat aiemmat kokeelliset tutkimukset modaliteettien käytöstä ja käyttäjien poikkeavista vuorovaikutustavoista. Toteutettujen fuusiomenetelmien tarkastelu puolestaan auttaa näkemään niissä olevat virheet ja kehitysmahdollisuudet, jolloin eri tekniikoita yhdistelemällä voidaan saada aikaan yhä tehokkaampia ratkaisuja.

Tutkielmassa käsitellyillä asioilla on siis merkitystä tulevia multimodaalisia järjestelmiä suunniteltaessa ja kehitettäessä. Mitä paremmin nämä asiat osataan ottaa huomioon, sitä helpompi on kehittää entistä toimintavarmempia ja joustavampia järjestelmiä. Tulevaisuudessa multimodaalisuus pystyy perinteisiä käyttöliittymiä paremmin vastaamaan uusien tietokonelaitteiden vaatimuksiin ja erilaisten vuorovaikutuskanavien kautta mahdollistamaan tietokoneen käytön yhä useammalle. Näin se täyttää vaatimuksen tietojen saavutettavuudesta tai toisin sanoen *design for all* -periaatteesta, jonka mukaan digitaalisen informaation on oltava kaikkien saatavilla. W3C osaltaan tukee tätä tarjoamalla EMMA-merkintäkielen edistämään uusien modaliteettien käyttöönottoa Internetin selailussa ja sen palveluissa. Nähtäväksi jää, miten nopeasti tai hitaasti multimodaalisuus tulee osaksi tavallisen käyttäjän jokapäiväistä elämää, ja onnistuuko EMMA koskaan saavuttamaan suosiota multimodaalisissa järjestelmissä.

8. Yhteenveto

Multimodaaliset järjestelmät muuttavat ihmisen ja tietokoneen välistä vuorovaikutusta kohti ihmisten luonnollista kommunikaatiota. Käyttäjien ei enää tarvitse tyytyä ainoastaan perinteisiin modaliteetteihin (hiiri, näppäimistö), vaan he voivat valita itselleen ja käyttötilanteeseensa parhaiten sopivan modaliteetin. Multimodaalisuus antaa käyttäjille useita keinoja ilmaista itseään, lisää käytön joustavuutta ja soveltuu käytettäväksi markkinoiden uusissa tietokonelaitteissa, joista monet pienen kokonsa vuoksi edellyttävät uusia käyttötapoja perinteisten rinnalle. Lisäksi se helpottaa tietokoneiden käyttöä, jolloin yhä useammat erityisryhmät voivat hyötyä tietotekniikasta. Huolimatta etuasemastaan perinteisiin käyttöliittymiin nähden ovat multimodaaliset järjestelmät vielä lapsenkengissään ja pitkälti tavallisen käyttäjän saavuttamattomissa.

Yksi multimodaalisille järjestelmille ominainen ongelma on käyttäjän eri modaliteeteilla antamien syötteiden yhdistäminen keskenään. Ennen varsinaista fuusioprosessia huomioon tulee ottaa useita asioita liittyen modaliteetteihin, syötteisiin ja itse toteutustekniikoihin. Käyttäjien väliset suuret yksilölliset erot modaliteettien käytössä hankaloittavat syötteiden käsittelyä, joskin yksilön kerran omaksuma multimodaalinen vuorovaikutustapa taasen on hyvin pysyvä ja vaikeasti muutettavissa. Käyttäjät valitsevat ja yhdistelevät modaliteetteja haluamallaan tavalla, vaikka käyttötilanteella ja suoritettavalla tehtävällä onkin vaikutusta heidän valintoihinsa.

Käytännössä käyttäjän omaksuma vuorovaikutustapa melko pian antaa viitteitä siitä, millaiseksi kommunikaatio tietokoneen kanssa muodostuu. Se, suosiiko käyttäjä samanaikaista vai jaksoittaista syötteiden antamista, vaikuttaa suoraan siihen, onko syötteiden välillä viive vai ovatko ne osittain päällekkäisiä. Syötteiden antaminen samanaikaisesti ei ole niin tavallista, kuin voisi multimodaalisten järjestelmien kohdalla odottaa, vaan useimmat käyttäjät antavat syötteet erillisinä, jolloin myös niiden väliset viiveajat vaihtelevat käyttäjäkohtaisesti. Syötteiden tulkinnan onnistumisen kannalta tärkeää onkin, että järjestelmä pystyy erottelemaan ja tulkitsemaan oikein uni- ja multimodaaliset syötteet.

Syötteiden oikeaa tulkintaa vaikeuttavat sekä niiden synkronisaatioon että tunnistamiseen liittyvät ongelmat. Multimodaalisen järjestelmän on sekä nopeasti annettava palaute käyttäjän syötteeseen että odotettava mahdollisia samaan syötekokonaisuuteen kuuluvia syötteitä. Odotusongelma onkin yritetty ratkaista määrittelemällä erilaisia aikakynnyksiä, joiden puitteissa annetut syötteet tulee yhdistää keskenään käyttäjän todellisen tavoitteen selvittämiseksi. Syötteiden tunnistaminen ei varsinaisesti ole multimodaalisuudesta johtuva ongelma, vaan koskee yleisesti tunnistamistekniikoita, jotka jo luonnostaan ovat virheherkkiä. Itse asiassa multimodaaliset järjestelmät osaltaan helpottavat tunnistamisongelmien ratkaisua tarjoamalla mahdollisuuden modaliteettien keskinäiseen kompensaatioon.

Vaikka syötteiden tulkinta ja yhdistäminen on yksi suurimmista haasteista multimodaalisia järjestelmiä suunniteltaessa, on niiden ratkaisemiseksi esitetty useita lupaavia toteutuksia. Eri modaliteeteilta tulleet syötteet yhdistää fuusioprosessi, jonka toteutustekniikka ja suoritustaso vaihtelevat valitusta fuusiomenetelmästä riippuen. Tavoitteena on saada aikaan yksiselitteinen yhteistulkinta erillisistä syötteistä, jonka perusteella tietokone voi antaa tarkoituksenmukaisen palautteen. Syötteiden yhdistämistä voidaan avustaa merkintäkieli EMMAa käyttämällä, joka on W3C:n kehittämä tiedonsiirtoformaatti. Tulevaisuudessa EMMAn on tarkoitus toimia virallisena standardina useita syötteitä tulkitseville järjestelmille, joissa se välittää informaatiota syöteprosessorien ja vuorovaikutuksenhallinnan välillä. Merkintäkieli tarjoaa useita elementtejä ja annotaatiota tulkintojen tekemistä varten ja saattaa olla yksi merkittävimmistä ratkaisuista multimodaalisten syötteiden käsittelyyn edellyttäen, että se yleisesti hyväksytään osaksi multimodaalisten järjestelmien sovelluskehitystä.

Viiteluettelo

[Adelhardt *et al*., 2003] Johann Adelhardt, Rui P. Shi, Carmen Frank, Viktor Zeissler, Anton Batliner, Elmar Nöth and Heinrich Niemann, Multimodal user state recognition in a modern dialogue system. In: *Proc. of the KI 2003: Advances in Artificial Intelligence, 26th Annual German Conference on AI, Lecture Notes in Computer Science* **2821** (2003), Springer, 591-605. Also available as *http://www.smartkom.org/Vortraege/ki2003.pdf*.

[Althoff *et al*., 2002] Frank Althoff, Marc Al-Hames, Gregor McGlaun and Manfred Lang, Towards a new approach for integrating multimodal user input based on evolutionary computation. In: *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing* **2** (2002), 2033-2036.

[Ao *et al*., 2007] Xiang Ao, Xugang Wang, Feng Tian, Guozhong Dai and Hongan Wang, Crossmodal error correction of continuous handwriting recognition by speech. In: *Proc. of the 12th International Conference on Intelligent User Interfaces* (2007), ACM Press, 243-250.

[Bernsen, 2002] Niels Ole Bernsen, Multimodality in language and speech systems – from theory to design support tool. In: Björn Granström, David House and Inger Karlsson (eds.), *Multimodality in Language and Speech Systems*. Text, Speech and Language Technology **19**, Kluwer Academic, 2002, 93-148. Also available as *http://www.nis.sdu.dk/demos/multimodality/multimodality.pdf*.

[Bolt, 1980] Richard A. Bolt, "Put-that-there": voice and gesture at the graphics interface. In: *Proc. of the 7th Annual Conference on Computer Graphics and Interactive Techniques* (1980), ACM Press, 262-270.

[Bunt, 1998] Harry Bunt, Issues in multimodal human-computer communication. In: *Proc. of Multimodal Human-Computer Communication: Systems, Techniques and Experiments, Lecture Notes in Computer Science* **1374** (1998), Springer, 1-12.

[Bunt *el al*., 2003] Harry Bunt, Michael Kipp, Mark T. Maybury and Wolfgang Wahlster, Fusion and coordination for multimodal interactive information presentation. In: Oliviero Stock and Massimo Zancanaro (eds*.), Multimodal Intelligent Information Presentation.* Text, Speech and Language Technology **27**, Kluwer Academic, 2003, 325-340. Also available as *http://www.dfki.de/~wahlster/Publications/Fusion_and_Coordination_for_Multimodal_Interactive_Information_Presentation.pdf*.

[Chai, 2002] Joyce Chai, Semantics-based representation for multimodal interpretation in conversational systems. In: *Proc. of the 19th International Conference on Computational Linguistics* (2002), ACL Press, 1-7.

[Chai *et al*., 2004] Joyce Y. Chai, Pengyu Hong and Michelle X. Zhou, A probabilistic approach to reference resolution in multimodal user interfaces. In: *Proc. of the 9th*

International Conference on Intelligent User Interfaces (2004), ACM Press, 70-77.

[Corradini *et al*., 2003] Andrea Corradini, Manish Mehta, Niels Ole Bernsen, Jean-Claude Martin and Sarkis Abrilian, Multimodal input fusion in human-computer interaction on the example of the NICE project. In: *Proc. of the NATO-ASI Conference on Data Fusion for Situation Monitoring, Incident Detection, Alert and Response Management* (2003). Available as *http://www.nis.sdu.dk/publications/2003/NATO-ASI_Armenia.pdf*.

[De Angeli *et al.*, 1998] Antonella De Angeli, Walter Gerbino, Giulia Cassano and Daniela Petrelli, Visual display, pointing, and natural language: the power of multimodal interaction. In: *Proc. of the Working Conference on Advanced Visual Interfaces* (1998*)*, ACM Press, 164-173.

[Flippo *et al*., 2003] Frans Flippo, Allen Krebs and Ivan Marsic, A framework for rapid development of multimodal interfaces. In: *Proc. of the 5th International Conference on Multimodal Interfaces* (2003), ACM Press, 109-116. Also available as *http://www.caip.rutgers.edu/disciple/Publications/icmi2003.pdf*.

[Gupta, 2003a] Anurag Gupta, An adaptive approach to collecting multimodal input. In: *Proc. of the 41st Annual Meeting on Association for Computational Linguistics* **2** (2003), ACL Press, 31-36. Also available as *http://acl.ldc.upenn.edu/P/P03/P03-2005.pdf*.

[Gupta, 2003b] Anurag Gupta, A reference model for multimodal input interpretation. In: *CHI '03 Extended Abstracts on Human Factors in Computing Systems* (2003), ACM Press, 936-937.

[Holzapfel *et al*., 2004] Hartwig Holzapfel, Kai Nickel and Rainer Stiefelhagen, Implementation and evaluation of a constraint-based multimodal fusion system for speech and 3D pointing gestures. In: *Proc. of the 6th International Conference on Multimodal Interfaces* (2004), ACM Press, 175-182. Also available as *http://isl.ira.uka.de/fileadmin/publication-files/holzapfel_icmi2004.pdf*.

[Huang and Oviatt, 2006] Xiao Huang and Sharon Oviatt, Toward adaptive information fusion in multimodal systems. In: *Proc. of Machine Learning for Multimodal Interaction*, *Lecture Notes in Computer Science* **3869** (2006), Springer, 15-27.

[Huang *et al*., 2006] Xiao Huang, Sharon Oviatt and Rebecca Lunsford, Combining user modeling and machine learning to predict users' multimodal integration patterns. In: *Proc. of Machine Learning for Multimodal Interaction*, *Lecture Notes in Computer Science* **4299** (2006), Springer, 50-62.

[Hurtig and Jokinen, 2006] Topi Hurtig and Kristiina Jokinen, Modality fusion in a route navigation system. In: *Proc. of the Workshop on Effective Multimodal Dialogue Interfaces* (2006). Available as

http://www.cs.uta.fi/research/hci/spi/MUMS/papers/EMMDI2006_hurtig_jokinen.pdf.

[Irawati *et al*., 2006] Sylvia Irawati, Scott Green, Mark Billinghurst, Andreas Duenser and Heedong Ko, An evaluation of an augmented reality multimodal interface using speech and paddle gestures. In: *Proc. of Advances in Artificial Reality and Tele-Existence, Lecture Notes in Computer Science* **4282** (2006), Springer, 272-283.

[Johnston, 1998] Michael Johnston, Unification-based multimodal parsing. In: *Proc. of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics* (1998), ACL Press, 624-630.

[Johnston and Bangalore, 2000] Michael Johnston and Srinivas Bangalore, Finite-state multimodal parsing and understanding. In: *Proc. of the 18th Conference on Computational Linguistics* (2000), ACL Press, 369-375. Available also as *http://www.research.att.com/~johnston/papers/colingmmfst.pdf*.

[Johnston *et al*., 1997] Michael Johnston, Philip R. Cohen, David McGee, Sharon L. Oviatt, James A. Pittman and Ira Smith, Unification-based multimodal integration. In: *Proc. of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics* (1997), ACL Press, 281-288.

[Johnston *et al*., 2007] Michael Johnston, Paolo Baggia, Jerry Carter, Deborah A. Dahl, Gerry McCobb and Dave Raggett, EMMA: Extensible MultiModal Annotation markup language. W3C Working Draft, 9 April 2007. Available as *http://www.w3.org/TR/emma/*.

[Kanninen, 2003] Matti Kanninen, Multimodaalisuus käyttöliittymäsuunnittelijan näkökulmasta - Graafisen ja puhekäyttöliittymän symbioosi. Taideteollinen korkeakoulu, Medialaboratorio, Lopputyö, huhtikuu 2003. Saatavilla myös *http://mlab.uiah.fi/www/projects_and_publications/final_thesis/pdf/kanninen_loppputyo*.

[Kaur *et al*., 2003] Manpreet Kaur, Marilyn Tremaine, Ning Huang, Joseph Wilder, Zoran Gacovski, Frans Flippo and Chandra Sekhar Mantravadi, Where is "it"? Event synchronization in gaze-speech input systems. In: *Proc. of the 5th International Conference on Multimodal Interfaces* (2003), ACM Press, 151-158.

[Käster *et al*., 2003] Thomas Käster, Michael Pfeiffer and Christian Bauckhage, Combining speech and haptics for intuitive and efficient navigation through image databases. In: *Proc. of the 5th International Conference on Multimodal Interfaces* (2003), ACM Press, 180-187.

[Larson, 2005] James A. Larson, Standard languages for developing multimodal applications. In: *Proc. of the 11th International Conference on Human-Computer Interaction* (2005). Available as *http://www.larson-tech.com/Writings/multimodal.pdf*.

[Larson *et al*., 2003] James A. Larson, T.V. Raman and Dave Raggett, W3C Multimodal Interaction Framework. W3C Note, 6 May 2003. Available as *http://www.w3.org/TR/mmi-framework/*.

[Lee and Yeo, 2005] Bee-Wah Lee and Alvin W. Yeo, Integrating sketch and speech inputs using spatial information. In: *Proc. of the 7th International Conference on Multimodal Interfaces* (2005), ACM Press, 2-9.

[Li *et al*., 2006] Li Li, Wu Chou, Feng Liu and Fei Cao, XM-flow: an Extensible Micro-flow for multimodal interaction. In: *Proc. of the 8th IEEE Workshop on Multimedia Signal Processing* (2006), 497-500. Available as *http://research.microsoft.com/workshops/mmsp06/MMSP2006_files/cameraReadyPapersNew/CameraReady_306.pdf*.

[MacLean, 2000] Karon E. MacLean, Designing with haptic feedback. In: *Proc. of the IEEE Robotics and Automation* (2000), 783-788. Available as *http://www.cs.ubc.ca/~maclean/publics/icra00-DesignWithHaptic-reprint.PDF*.

[McKenzie Mills and Alty, 1998] Karen McKenzie Mills and James L. Alty, Investigating the role of redundancy in multimodal input systems. In: *Proc. of the International Gesture Workshop on Gesture and Sign Language in Human-Computer Interaction, Lecture Notes in Computer Science* **1371** (1998), Springer, 159-171.

[Malmberg, 2006] Tanja Malmberg, Standardit multimodaalisissa käyttöliittymissä. Roope Raisamo (toim.), Käyttöliittymien ohjelmistoarkkitehtuurit, Raportti **B-2006-1**, Tietojenkäsittelytieteiden laitos, Tampereen yliopisto, 2006, 45-61.

[Mankoff and Abowd, 1999] Jennifer C. Mankoff and Gregory D. Abowd, Error correction techniques for handwriting, speech, and other ambiguous or error prone systems. Georgia Institute of Technology, Technical Report **GIT-GVU-99-18**, February 1999. Available as *http://smartech.gatech.edu/bitstream/1853/3385/1/99-18.pdf*.

[Martin *et al*., 1998] Jean-Claude Martin, R. Veldman and Dominique Béroule, Developing multimodal interfaces: a theoretical framework and guided propagation networks. In: *Proc. of Multimodal Human-Computer Communication: Systems, Techniques, and Experiments, Lecture Notes in Computer Science* **1374** (1998), Springer, 158-187.

[Maybury, 2002] Mark T. Maybury, Language technology – a survey of the state of the art: language resources – multimodal language resources. MITRE Corporation, Technical Paper, December 2002. Available as

http://www.mitre.org/work/tech_papers/tech_papers_02/maybury_language/maybury_language_tech.pdf.

[Maybury and Wahlster, 1998] Mark T. Maybury and Wolfgang Wahlster, Intelligent user interfaces: an introduction. In: Mark T. Maybury and Wolfgang Wahlster (eds.), *Readings in Intelligent User Interfaces*. Morgan Kaufmann, 1998, 1-14.

[Milota, 2004] André D. Milota, Modality fusion for graphic design applications. In: *Proc. of the 6th International Conference on Multimodal Interfaces* (2004), ACM Press, 167-174.

[Moran *et al*., 1997] Douglas B. Moran, Adam J. Cheyer, Luc E. Julia, David L. Martin and Sangkyu Park, Multimodal user interfaces in the Open Agent Architecture. In: *Proc. of the 2nd International Conference on Intelligent User Interfaces* (1997), ACM Press, 61-68.

[Nigay and Coutaz, 1993] Laurence Nigay and Joëlle Coutaz, A design space for multimodal systems: concurrent processing and data fusion. In: *Proc. of the SIGCHI Conference on Human Factors in Computing Systems* (1993), ACM Press, 172-178.

[Nigay and Coutaz, 1995] Laurence Nigay and Joëlle Coutaz, A generic platform for addressing the multimodal challenge. In: *Proc. of the SIGCHI Conference on Human Factors in Computing Systems* (1995), ACM Press, 98-105. Also available as *http://acm.org/sigchi/chi95/proceedings/papers/lmn_bdy.htm*.

[Oviatt, 1999a] Sharon Oviatt, Ten myths of multimodal interaction. *Communications of the ACM* **42**, 11 (Nov. 1999), 74-81.

[Oviatt, 1999b] Sharon Oviatt, Mutual disambiguation of recognition errors in a multimodal architecture. In: *Proc. of the SIGCHI Conference on Human Factors in Computing Systems: the CHI is the limit* (1999), ACM Press, 576-583.

[Oviatt, 2000] Sharon Oviatt, Taming recognition errors with a multimodal interface. *Communications of the ACM* **43**, 9 (Sep. 2000), 45-51.

[Oviatt and Lunsford, 2005] Sharon Oviatt and Rebecca Lunsford, Multimodal interfaces for cell phones and mobile technology. *International Journal of Speech Technology* **8**, 2 (Jun. 2005), 127-132.

[Oviatt and Olsen, 1994] Sharon Oviatt and Erik Olsen, Integration themes in multimodal human-computer interaction. In: *Proc. of the International Conference on Spoken Language Processing* **2** (1994), Acoustical Society of Japan, 551-554.

[Oviatt *et al.*, 1997] Sharon Oviatt, Antonella DeAngeli and Karen Kuhn, Integration and synchronization of input modes during multimodal human-computer interaction. In: *Proc. of the SIGCHI Conference on Human Factors in Computing Systems* (1997), ACM Press, 415-422.

[Oviatt *et al.*, 2000] Sharon Oviatt, Phil Cohen, Lizhong Wu, John Vergo, Lisbeth Duncan, Bernhard Suhm, Josh Bers, Thomas Holzman, Terry Winograd, James Lan-

day, Jim Larson and David Ferro, Designing the user interface for multimodal speech and pen-based gesture applications: state-of-the-art systems and future research directions. *Human-Computer Interaction* **15**, 4 (Aug. 2000), 263-322.

[Oviatt *et al*., 2003] Sharon Oviatt, Rachel Coulston, Stefanie Tomko, Benfang Xiao, Rebecca Lunsford, Matt Wesson and Lesley Carmichael, Toward a theory of organized multimodal integration patterns during human-computer interaction. In: *Proc. of the 5th International Conference on Multimodal Interfaces* (2003), ACM Press, 44-51.

[Oviatt *et al*., 2005a] Sharon Oviatt, Rebecca Lunsford and Rachel Coulston, Individual differences in multimodal integration patterns: what are they and why do they exist? In: *Proc. of the SIGCHI Conference on Human Factors in Computing Systems* (2005), ACM Press, 241-249. Also available as *http://www.cse.ogi.edu/CHCC/Groups/ReadingGroup/articles/chi05.pdf*.

[Oviatt *et al*., 2005b] Sharon Oviatt, Rachel Coulston and Rebecca Lunsford, Just do what I tell you: the limited impact of instructions on multimodal integration patterns. In: *Proc. of User Modeling'05, Lecture Notes in Computer Science* **3538** (2005), Springer, 261-270.

[Paleari and Lisetti, 2006] Marco Paleari and Christine L. Lisetti, Toward multimodal fusion of affective cues. In: *Proc. of the 1st ACM International Workshop on Human-centered Multimedia* (2006), ACM Press, 99-108.

[Paternò, 2004] Fabio Paternò, Multimodality and multi-device interfaces. In: *W3C Workshop on Multimodal Interaction* (2004). Available as *http://www.w3.org/2004/02/mmi-workshop/paterno-cnr.pdf*.

[Pfleger, 2004] Norbert Pfleger, Context based multimodal fusion. In: *Proc. of the 6th International Conference on Multimodal Interfaces* (2004), ACM Press, 265-272.

[Portillo *et al*., 2006] Pilar Manchón Portillo, Guillermo Pérez García and Gabriel Amores Carredano, Multimodal fusion: a new hybrid strategy for dialogue systems. In: *Proc. of the 8th International Conference on Multimodal Interfaces* (2006), ACM Press, 357-363.

[Reithinger *et al*., 2005] Norbert Reithinger, Simon Bergweiler, Ralf Engel, Gerd Herzog, Norbert Pfleger, Massimo Romanelli and Daniel Sonntag, A look under the hood – design and development of the first SmartWeb system demonstrator. In: *Proc. of the 7th International Conference on Multimodal Interfaces* (2005), ACM Press, 159-166. Also available as *http://www.dfki.de/~flint/papers/icmi05.pdf*.

[Riding and Cheema, 1991] Richard Riding and Indra Cheema, Cognitive styles - an overview and integration. *Educational Psychology* **11**, 3/4 (1991), 193-215.

[Russ *et al*., 2005] Gerhard Russ, Brian Sallans and Harald Hareter, Semantic based information fusion in a multimodal interface. In: *Proc. of the International Conference on Human-Computer Interaction* (2005), 94-100. Available as *http://members.chello.at/hoebertz-sallans/sallans/papers/MMI_Fusion_2005.pdf*.

[Salber *et al*., 1995] Daniel Salber, Joëlle Coutaz, Laurence Nigay (eds.), Giorgio Faconti, Fabio Paternò, David Duke and Michael Harrison, The system modelling glossary. European ESPRIT Project Amodeus-2, Technical Report **SM/WP 26**, June 1995. Available as *http://daniel.salber.name/publications/amodeus_sm_wp26.pdf*.

[Streit, 2001] Michael Streit, Why are multimodal systems so difficult to build? - About the difference between deictic gestures and direct manipulation. In: *Revised Papers from the Second International Conference on Cooperative Multimodal Communication, Lecture Notes in Computer Science* **2155** (2001), Springer, 176-196.

[Sun *et al*., 2006a] Yong Sun, Fang Chen and Vera Chung, QuickFusion: multimodal fusion without time thresholds. In: *Proc. of the 2005 NICTA-HCSNet Multimodal User Interaction Workshop* **57** (2006), Australian Computer Society, 51-54.

[Sun *et al*., 2006b] Yong Sun, Fang Chen, Yu (David) Shi and Vera Chung, A novel method for multi-sensory data fusion in multimodal human computer interaction. In: *Proc. of the 20th Conference of the CHISIG of Australia on Computer-Human Interaction: design, activities, artefacts and environments* (2006), ACM Press, 401-404.

[Vo and Waibel, 1997] Minh Tue Vo and Alex Waibel, Modeling and interpreting multimodal inputs: a semantic integration approach. Carnegie Mellon University, Technical Report **CMU-CS-97-192**, December 1997. Available as *http://reports-archive.adm.cs.cmu.edu/anon/1997/CMUCS-97-192.ps*.

[West *et al*., 2004] David West, Trent Apted and Aaron Quigley, A context inference and multi-modal approach to mobile information access. In: *Proc. of the Artificial Intelligence in Mobile Systems* (2004), 28-35. Also available as *http://w5.cs.uni-sb.de/~baus/aims04/cameraready/P5.pdf*.

[Wu *et al*., 2002] Lizhong Wu, Sharon L. Oviatt and Philip R. Cohen, From members to teams to committee – a robust approach to gestural and multimodal recognition. *IEEE Transactions on Neural Networks* **13**, 4 (Jul. 2002), 972-982.

[Xiao *et al*., 2000] Bin Xiao, Jiantao Pu and Shihai Dong, Multimodal integration using complex feature set. In: *Proc. of the 3rd International Conference on Advances in Multimodal Interfaces, Lecture Notes in Computer Science* **1948** (2000), Springer, 245-252.

[Zhang *et al*., 2004] Qiaohui Zhang, Atsumi Imamiya, Kentaro Go and Xiaoyang Mao, Resolving ambiguities of a gaze and speech interface. In: *Proc. of the Symposium on Eye Tracking Research & Applications* (2004), ACM Press, 85-92.