Datoranimerade talande ansikten

Olov Engwall, Centrum för talteknologi, Kungliga Tekniska Högskolan, Stockholm
Videoillustrationer (markerade med VI i texten) till detta kapitel finns på http://www.speech.kth.se/~olov/ansiktemotansikte.html.

Kommunikation mellan människor och datorer kan underlättas av att datorn representeras av en virtuell personlighet, ett datoranimerat ansikte. Som datoranvändare kan man samtala med ett sådant ansikte, ställa frågor och få svar, på samma sätt som när man talar med en verklig människa. Därigenom kan interaktionen med datorn följa invanda sätt att kommunicera. Animerade ansikten kan också underlätta samtal mellan människor genom att vara en stödjande länk för exempelvis läppavläsning. Detta kapitel beskriver forskning kring datoranimerade talade ansikten: hur de kan användas och hur man med modeller och mätningar försöker göra ansiktsrörelser och uttryck så lättolkade eller naturliga som möjligt. Avslutningsvis diskuteras hur realistiska de datoranimerade ansiktena kan och bör göras.

Varför ska man använda datoranimerade talande ansikten?

I kommunikation mellan människor är icke-verbala signaler mycket viktiga. Ansiktsuttryck, blick, gester och röstkvalité står för en stor del av den överförda informationen. Ofta är denna del till och med större än den som ges av orden i sig. Hur viktiga de icke-verbala signalerna är varierar beroende på vilken typ av kommunikation det gäller. För att uttrycka gillande eller ogillande fann Mehrabian i en känd studie från 1971 att orden inte stod för mer än 7 %, medan röstkvalité stod för 38 % och kroppsspråk för 55 %. Mehrabians studie har ofta felaktigt generaliserats till att fördelningen skulle gälla för all audiovisuell kommunikation. Med en sådan generalisering underskattar man ordens betydelse i allmänhet, men sociologer som Birdwhistell och Hall hävdar att orden överför högst 30-40 % av informationen i social interaktion och att resten är icke-verbal. Genom att lägga till ett virtuellt ansikte kan därför kommunikationen mellan användare och datorn inte bara bli naturligare och rikare, utan också effektivare. Detta eftersom vår perception använder sig av flera olika kanaler och vi uppfattar vad som sägs utifrån både vad vi hör och vad vi ser. Informationen från hörseln och synen kompletterar varandra. Därför är det mycket lättare att höra vad en person säger i en bullrig miljö, på exempelvis en restaurang, om man samtidigt ser personens ansikte. En studie från Kungliga Tekniska Högskolan (KTH) i

Stockholm visar exempelvis att personer med normal hörsel korrekt uppfattar nästan dubbelt så många vokal-konsonant-vokal-sekvenser (exempelvis ”appa”) som spelas upp i högt brus om de samtidigt får se talarens ansikte (58 % mot 30 % korrekta ord om de enbart får höra ljudet). Med ett datoranimerat ansikte blev resultatet nästan lika bra som med det naturliga (55 % korrekta ord). Motsvarande förbättring kan uppnås för hörselskadade personer som lyssnar på ord utan brus.

Genom att använda ett datoranimerat ansikte kan man inte bara göra signaler som finns i de talade orden tydligare, såsom ljud och betoning, utan också ge sådan information som inte sägs verbalt, som blick och gester. Dessutom kan ansiktet uttrycka känslor och därmed förstärka eller förändra innebörden av det som sägs. Tolkningen av en mening som ”Han kommer vara där” blir helt annorlunda beroende på om den ackompanjeras av ett glatt eller ett ledset ansiktsuttryck, särskilt om tonfallet i sig är neutralt. Ett animerat ansikte är alltså användbart genom att det kan både förstärka och komplettera den givna informationen.

En ytterligare dimension som framkommit i olika forskningsstudier är att användaren ofta blir mer nöjd med kvalitén på interaktionen och på datorsystemet om den sker med ett talande ansikte, även om informationen användaren får är densamma. Att interagera med en personifierad gestalt, snarare än med ett anonymt system, påverkar alltså upplevelsen.

Vad är ett datoranimerat talande ansikte?

Den första associationen till datoranimerade talande ansikten är kanske biograffilmer, och då främst de många Hollywood-produktioner av bland andra Pixar och DreamWorks, där alla filmens skådespelare är datoranimerade karaktärer. Vid stöd för talad kommunikation handlar det i likhet med dessa filmer om att skapa nya rörelser och uttryck hos en karaktär med hjälp av datorprogram. Den stora skillnaden är att datoranimerade ansikten i filmer i hög utsträckning bearbetas bild för bild av en stab med animatörer, medan målet för forskningen är att animationen ska ske automatiskt, från en skriven text eller ett inspelat yttrande. På så sätt kan ansiktet agera självständigt och uttrycka även sådant som inte tidigare skapats av en mänsklig animatör.

Även uttrycket ”talande ansikten” är värt att diskutera. Eftersom ansiktena skapas för att främja talad kommunikation mellan människor och datorn är det något missvisande att enbart tala om ansikten. På engelska använder man istället termen ”Talking Heads” (talande huvuden), vilket inte enbart är en blinkning till det gamla rockbandet med samma namn. Termen poängterar också att det handlar om mer än bara ett ansikte. Även huvudets rörelser och hur det är vridet ger viktig information. Ännu viktigare är att huvudet har ett innehåll och inte bara ansiktets yta. Ett huvud kan tänkas ha en tolkande ”hjärna” inuti, men framför allt hat det en mun bakom läpparna. Om och hur vi ser talarens tänder och tungspets har stor inverkan på hur vi förstår det som sägs. Därför är det viktigt att dessa delar finns med.

Med ett ”talande ansikte” menas alltså i första hand ett virtuellt huvud som har ett ansikte, tunga och tänder. Dessutom att ansiktsuttryck och rörelser är gjorda för att vara så lättolkade som möjligt när ansiktet talar.

Hur använder man datoranimerade talande ansikten?

Inledningsvis beskrevs hur datoranimerade talande ansikten kan förbättra människa-datorinteraktion genom att göra kommunikationen mer naturlig. Forskning inom detta område, kallat multimodala dialogsystem, bedrivs sedan mer än ett decennium vid KTH. Den kretsar kring olika tekniker för att tolka vad användaren säger, för att generera talade svar från datorn, för att göra dialogen så naturlig som möjlig och för att skapa talande ansikten som stödjer interaktionen. Syftet med multimodala dialogsystem är oftast att användaren ska kunna få upplysningar från en databas med information om exempelvis tågtidtabeller, telefonnummer eller restauranger. Liknande system finns redan som automatiska telefonupplysningstjänster, där användaren ringer och får informationen uppläst i luren. Skillnaden mellan dessa telefonupplysningstjänster och ett multimodalt dialogsystem ligger både i hur olika kommunikationskanaler, modaliteter, utnyttjas och i hur fri dialogen är.

I ett multimodalt gränssnitt kan användaren överföra den information datorn behöver på flera olika sätt, framför allt med sin röst och genom att peka med datormusen. Även datorn kan presentera informationen både muntligt och visuellt, i tabeller eller i kartor, beroende

på vad som är lämpligt för just den typ av information som den ska ge. Ett talande ansikte kan fylla en viktig funktion i ett sådant dialogsystem, genom att vara den virtuella person, ofta kallad agent, som visar upp de olika typerna av information för användaren.

I dagens automatiska telefonupplysningstjänster är dialogen låst till ett förutbestämt mönster. Systemet frågar stegvis den som ringer efter information och därefter läses den önskade upplysningen upp. Syftet med forskningen kring dialogsystem är att användaren ska kunna ta mer initiativ i dialogen. Därför är det som kallas turtagning viktigt. Det innebär att samtalsparterna med, oftast omedvetna, signaler visar när de själva vill säga något, när de lämnar över ordet, och om de följer med i det den andra personen säger. I informationstjänster där systemet bestämmer när användaren får prata behövs ingen turtagning, men ett friare dialogsystem måste kunna ge signaler om vems tur det är att tala.

Dialogsystemet AdApt är ett tydligt exempel på hur ett talande ansikte kan användas för att förenkla både informationsutbyte och turtagning mellan människor och dator. Användaren och det talande ansiktet resonerar sig tillsammans fram till lämpliga lediga lägenheter i Stockholm utifrån användarens krav på område, storlek, pris och speciella önskemål om exempelvis balkong eller öppen spis. Lägenheterna visas som färgade ikoner på en karta, som i Figur 1, och agentens blick och huvudrörelser visar var på kartan användaren ska titta. Agenten ger också signaler, med i första hand ansiktets rörelser, som underlättar turtagningen. När användaren talar, visar agentens uttryck att han lyssnar uppmärksamt. Om användaren tystnar och systemet förväntar sig en fortsättning, för att det som användaren sagt är ofullständigt, visar agenten det genom ett ansiktsuttryck som uppmuntrar användaren att fortsätta meningen. Om systemet istället får uppfattningen att frågan är avslutad och att det är dags att svara, visar agentens ansiktsuttryck att han förstått att det är hans tur i dialogen. Under tiden som systemet letar i databasen efter lägenheter som passar in på användarens krav ser agenten ut att tänka efter, för att på så sätt visa att frågan bearbetas. Därefter tittar agenten upp mot användaren för att svara och återgår sedan till ett uppmärksamt lyssnande ansiktsuttryck för att visa att han återlämnar initiativet till användaren. Genom dessa signaler om vems tur det är att tala och när systemet är upptaget

med att leta, förenklas interaktionen, eftersom användaren får information om situationen i dialogen, utan att det sägs rent ut.

[Figur 1 ungefär här]

Det talande ansiktet används också för att förstärka den information som agenten ger muntligt och på så sätt göra den tydligare. När man vill betona något ord i det man säger, kan det akustiskt göras längre, med mer ljudtryck och genom att förändra intonationen. Denna betoning kan förstärkas genom huvudets uttryck och rörelser så att det blir tydligare vad som är viktigt. Om användaren exempelvis frågar ”Har den gula lägenheten balkong?” och systemet svarar ”Nej, men den röda lägenheten har balkong” är det naturligt att ordet ”röda” betonas för att visa att det är den viktigaste informationen. Om ansiktet nickar lite eller höjer ögonbrynen samtidigt som ”röda” sägs, blir betoningen tydligare. Andra studier av hur försökspersoner som tittar på ett talande ansikte uppfattar betoning visar att nickning och ögonbrynsrörelse har en stark effekt på vilket ord som uppfattas som betonat. Betoningen uppfattas tydligast om båda rörelserna görs samtidigt, men nickning är den viktigaste signalen. På samma sätt kan en nickning, ett leende, höjda eller rynkade ögonbryn visa om det agenten säger är en bekräftelse, en fråga eller ett ifrågasättande. Precis samma yttrande kan därigenom få olika betydelse bara genom ansiktsuttrycket. Om användaren i AdApt säger ”Berätta om den röda lägenheten” och agenten svarar ”Den röda” och samtidigt nickar och ler betyder det ungefär ”Jag förstod att du frågade om den röda lägenheten. Ett ögonblick så ska jag leta upp information om den”. Med ett mindre leende och höjda ögonbryn kan man istället tolka det som ”Jag är osäker på att jag uppfattade rätt, var det den röda du sa?”. Utan leende, med rynkade ögonbryn och öppnare ögon kan budskapet istället vara ”Jag måste ha hört fel; det finns ju ingen röd lägenhet på kartan”. Denna typ av ofullständiga meningar, vars betydelse visas av intonationen och ansiktsuttrycket, är mycket vanlig när människor talar. Genom att utnyttja det i människa-datorinteraktion kan informationsutbytet gå snabbare, eftersom man undviker långa explicita bekräftelser eller frågor, som snabbt upplevs som tjatiga.

Som framgått av det första avsnittet, kan ett talande ansikte ge ett viktigt stöd för att uppfatta vad som sägs, antingen i en bullrig miljö, eller om åhöraren är hörselskadad. Detta

har utnyttjats i ett tidigare EU-projekt och nu av ett litet nystartat företag, båda med namnet Synface. Grundidén är att skapa en datoranimerad videotelefon för hörselskadade (VI 1). Systemet, Eyephone, riktar sig till dem som klarar av att föra ett samtal ansikte mot ansikte genom att kombinera ljudinformation med läppavläsning, men som har svårigheter vid telefonsamtal, eftersom de då inte kan se den som talar. I framtiden är det mycket möjligt att problemet lätt löses genom att bildtelefoner, där de som samtalar även kan se varandra, blir allmänt spridda. Det kräver dock hög bildkvalité och överföringshastighet för att möjliggöra läppavläsning och dithän har vi än så länge inte nått. Hittills har hörselskadade personer istället tvingats använda sig av en telefonförmedlingstjänst och speciell utrustning, så att samtalspartnern får prata med en telefonist som skriver ned det han eller hon säger och skickar över ett textmeddelande till den hörselskadade. Denna metod är naturligtvis krånglig, resurskrävande och integritetshämmande. Att kunna föra ett telefonsamtal direkt skulle därför vara en stor vinst för många hörselskadade och arbetet med Eyephone har också belönats med flera priser.

Den stora fördelen med Eyephone jämfört med videotelefoni är att den person som samtalar med den hörselskadade kan ha en helt vanlig telefon och att enbart ljudet behöver överföras. Istället har den hörselskadade ett speciellt datorprogram som tolkar det som samtalspartnern säger med automatisk taligenkänning. Denna tolkning skickas därefter till animeringsprogrammet, som använder tolkningen för att skapa motsvarande rörelser i det talande ansiktet. Det ursprungliga ljudet spelas slutligen upp tillsammans med ansiktets rörelser, så att den hörselskadade samtidigt kan läppavläsa och hämta information från ljudet. Eftersom taligenkänningen och animeringen tar lite tid att genomföra, fördröjs allt det samtalspartnern säger med 0,2 sekunder, men detta är inte längre tid än att samtalet kan föras på ett normalt sätt.

Ett alternativ till att använda taligenkänningens tolkning för att animera ett talande ansiktet skulle kunna vara att tolkningen presenteras direkt som en text på den hörselskadades datorskärm. Det finns dock två klara problem med den metoden. För det första talar vi inte som vi skriver och därför är det svårt att läsa en nedskriven talad text med upprepningar, stakningar, hummanden och orddelar som inte uttalas. För det andra är det talande ansiktet

betydligt mindre känsligt för fel i taligenkänningen, eftersom felen blir mindre tydliga i ansiktet. Det är därför betydligt lättare att korrekt tolka det som sägs med ett ansikte som gör några fel än om samma fel står skrivet. Med ett talande ansikte behöver igenkännaren inte heller känna igen hela ord, utan enbart ljuden. Som exempel ser orden ”stjäl”, ”skäl” och ”själ” likadana ut i ansiktet, och därför spelar det för Eyephone ingen roll vilket av orden som sagts, medan det däremot skulle vara förvirrande med en förväxling i skrift.

Hur bra Eyephone fungerar beror inte bara på taligenkänningen, utan också på individen. Typ och grad av hörselskada och hur van personen är att läsa på läpparna påverkar. För vissa är det talande ansiktet enbart något bättre än att bara få höra ljudet, och ibland kan det till och med göra det svårare, om taligenkänningen gjort något allvarligare misstag. För de flesta hörselskadade hjälper det talande ansiktet så att det är möjligt att förstå betydligt mer än utifrån enbart ljudet. Försök där meningar av typen ”Katten lekte med ett nystan” spelas upp visar att andelen ord som uppfattas korrekt med ett talande ansikte (61 %) är nästan dubbelt så stor som om enbart ljudet spelas upp (31 %), men inte riktigt lika hög som med både ljud och bild av den verkliga talaren (66 %). För ytterligare vissa kan det talande ansiktet ibland till och med vara bättre än det verkliga. Anledningen är att det är möjligt att överdriva rörelserna i det animerade ansiktet, så att de blir tydligare (VI 2).

Talande ansikten kan inte bara göra tydligare ansiktsrörelser än riktiga människor, det finns också möjlighet att låta dem göra sådant som är omöjligt för en mänsklig talare. Ett exempel är möjligheten att göra delar av huden genomskinlig för att visa hur munnen ser ut inuti för olika ljud. Det kan utnyttjas i virtuella språklärare eller talpedagoger, som kan hjälpa hörselskadade barn eller personer som lär sig ett främmande språk att förbättra sitt uttal. Bakgrunden är att både vuxna som lär sig ett nytt språk och hörselskadade barn kan ha svårt att höra skillnad på olika ljud. I det ena fallet kan det bero på att skillnaderna inte finns i det egna språket och i det andra på att hörseln är skadad i det frekvensintervall där skillnaden finns. Om man inte hör skillnaden, är det oerhört svårt att själv göra den. En datoranimerad lärare kan då hjälpa till genom att visa skillnaden i artikulationen.

I ARTUR-projektet utvecklas en sådan lärare, Artur. Han ska analysera elevens uttal och därefter ge återkoppling till eleven, om vad som ska förändras för att uttalet ska bli bättre (VI 3). Systemet består av flera olika delar. En komponent ska automatiskt hitta uttalsfel i det eleven säger, en annan återskapa hur eleven sade det ord som blev fel och en tredje generera instruktioner och animationer för att hjälpa eleven att komma till rätta med felet.

I dagsläget finns enbart prototyper av systemet, som testats under simulerade förhållanden. Responsen från användarna som i försöket skull lära sig hur det svenska ”sj”-ljudet ska uttalas har varit mycket positiv. Både svenska barn som går i träning hos talpedagoger och icke svensktalande vuxna som fått öva sig på svenska ord har varit entusiastiska över nyttan av att kunna se uttalet i det talande huvudet. Arbetet med ARTUR fick också en internationell utmärkelse 2004.

Att ”sj”-ljudet valdes ut i försöken beror på att en hörselskada ofta påverkar de ljudfrekvenser som är viktiga i ”sj” och på att det är ett unikt svenskt ljud, vilket gör det svårt för icke svensktalande. Dessutom är det svårt att från en vy av enbart ansiktet förstå hur ljudet uttalas, medan det blir tydligare om det är möjligt att se tungans rörelse. En viktig del av arbetet i ARTUR-projektet är hur dessa tungrörelser ska visas för användaren. Eftersom vi är vana vid att se ansikten och tolka dem, men däremot inte tungans rörelser, mer än det som syns i munöppningen, är det viktigt att animeringen av tungans rörelser görs lättbegriplig. Flera olika möjligheter finns, som att göra ansiktet genomskinligt eller att ta bort en del av kinden och visa tungans rörelser från sidan. De personer som fått pröva systemet berättar i intervjuer att det visserligen till en början kändes ovant att försöka överföra de tungrörelser som det talande ansiktet visade till den egna tungan, men att de genom bilderna relativt snart blev medvetna om hur de skulle försöka göra.

För att Artur ska fungera för uttalsträning och Eyephone för läppavläsning krävs att de ansikts- och tungrörelser som visas är både korrekta och lättförståeliga. De två följande avsnitten behandlar animationstekniker och mätningar som säkerställa detta.

Hur datoranimeras talande ansikten?

Animation av talande ansikten sker antingen med bearbetade videobilder av en verklig talare eller genom att en ren datormodell av ansiktet skapas.

I tekniken med bearbetade videobilder spelas en databas in med en person som läser upp ett textmaterial. Materialet ska innehålla ansiktsuttrycken för alla ljud man behöver kunna animera. Dessa ansiktsuttryck för ett visst ljud kallas *visem*, i analogi med beteckningen *fonem* för ett visst ljud. Eftersom visem skiljer sig åt beroende på både ljudet i sig och på de omgivande, måste alla visem spelas in i relevanta omgivningar. Exempelvis ser ”p” i ”sopor” helt annorlunda ut än i ”pippi”, eftersom det i det första fallet görs med framskjutna, rundade läppar och i det andra med mer tillbakadragna och raka. Å andra sidan finns det grupper av flera fonem som tillsammans utgör ett visem, eftersom de har identiska ansiktsuttryck, som [p, b, m] eller [t, d, n]. Inom varje grupp är ansiktsuttrycken utbytbara sinsemellan Därmed behöver man exempelvis inte spela in nya bilder för att kunna animera sekvensen ”Mimmi” om man redan spelat in ”pippi”.

När ett nytt yttrande ska animeras skarvas olika små videosekvenser ihop till den nya. Den stora svårigheten med denna metod är att undvika att den nya sekvensen hackar, beroende på att talarens ansiktsuttryck eller hållning ändrats mellan två bilder som sammanfogats. För att undvika sådana hopp finns flera metoder. En är att först välja ut nyckelbilder som innehåller de viktigaste visemen i sekvensen och därefter successivt smälta samman mellanliggande bilderna för att få en naturlig övergång, så kallad morfning (VI 4). För att kunna göra det krävs att man kan identifiera likheten mellan olika bilder och hitta lämpliga övergångar mellan dem. Även om det nu utvecklas tekniker för att göra detta automatiskt, krävs fortfarande oftast manuell märkning, vilket förstås är ett stort extra arbete. Som en förenkling kan man ta en av de ursprungliga inspelade sekvenserna där något annat sagts och enbart klistra in nya bilder med rätt munrörelser över munnen i ursprungsinspelningen. På så sätt är bilden i övrigt helt naturlig och istället för att hitta övergångar för hela ansiktet är det bara bilderna av munnen man behöver bearbeta. Ett stort problem med metoden är att det är osäkert om ansiktsuttrycket i övrigt passar ihop med munrörelserna. Ett alternativ är

därför att dela in ansiktet i videobilderna i flera olika regioner som sparas separat och sätts ihop till en ny helhet vid animeringen (VI 5).

Fördelarna med dessa videobaserade metoder är att bildkvalitén bitvis kan bli mycket bra och ansiktsrörelserna helt naturliga för sekvenser som finns i det inspelade materialet. Nackdelarna är dels att man kan få mycket onaturliga effekter när det inte finns några passande bilder i ursprungsinspelningarna, dels att ansiktet lätt kan uppfattas som stelt och känslokallt (se VI 4). För att kunna sätta ihop sekvenser som kommer från olika tidpunkter i inspelningsmaterialet måste talaren nämligen ha nästan samma kroppshållning och känsloläge i de bildsekvenser som klippas ihop. Det gör att man därför måste spela in hela databasen för varje känsloläge och kroppshållning som man behöver. Görligare, och vanligare, är istället att talaren under hela inspelningen har en mycket neutral hållning och känsla. För vissa användningsområden kan detta vara fullt acceptabelt, medan det för många andra är nödvändigt att kunna modifiera uttryck och hållning mer. Denna större frihet kan fås genom att skapa en datormodell av ansiktet.

Modellbaserad animering innebär att ansiktet beskrivs av ett tredimensionellt rutnät, såsom i Figur 2 och i VI 6. Rutnätet ändras för att skapa ansiktsrörelser genom *interpolering*, *direkt parametrisering*, *muskelmodeller* eller *datadrivna metoder*.

[Figur 2 ungefär här]

Interpolering är den enklaste och mest allmänt spridda metoden, eftersom den finns i de flesta kommersiella program för 3D-modellering och animering. Den har i grunden stora likheter med morfningstekniken för videobaserad animering, eftersom den utgår från att ett antal viktiga ansiktsformer skapas och sparas. Vid animeringen kan ansiktsformerna blandas, både i en enskild bild och mellan bilder som följer på varandra. Fördelen med metoden är att den är enkel, men det finns också nackdelar. En är att animeringen begränsas av vilka ansiktsformer som genererats och sparats, vilket gör att ett stort antal ansiktsformer först måste skapas. En annan nackdel är att interpolering mellan olika ansiktsformer inte

alltid ger naturliga övergångar, eftersom ansiktsrörelser ofta är komplicerade och normalt inte består av raka rörelser mellan ett läge och ett annat.

Direkt parametrisering föreslogs därför 1982 av Parke för att lösa problemen med interpoleringsmetoden. Istället för att skapa en katalog med ansiktsformer använde sig Parke av enkla geometriska förändringar av ansiktsmodellens rutnät. Punkterna i rutnätet flyttas genom att aktivera olika parametrar, närmast likt en marionetteater, som om animeringens regler drar i olika linor för att skapa olika rörelser. En parameter kontrollerade käkens rotation, en annan munnens bredd, en tredje blickpunkt och så vidare. Eftersom metoden är både intuitivt enkel och beräkningsmässigt effektiv har den fått stort genomslag och används som animeringsmetod för många animerade talande ansikten världen över. Metodens enkelhet är möjligen också dess svaghet, eftersom det inte finns någon garanti för att de enkla geometriska förändringarna motsvarar ansiktsrörelserna hos en människa som talar. Det problemet attackeras i huvudsak på två sätt: antingen försöker man få rörelserna naturliga genom modeller av ansiktsmusklerna eller genom att modellen får ”lära sig” hur ett riktigt ansikte rör sig. I det senare fallet gör man inspelningar av en eller flera personer för att på så sätt få fram hur deras ansikte ser ut när de talar.

Muskelbaserad animering innebär att man antar att varje punkt i nätet som bygger upp ansiktet har en tyngd och är ihopkopplad med omkringliggande punkter i ett nätverk av fjädrar. Beroende på hur tunga punkterna är, hur styva fjädrarna är och hur mycket man drar i dem skapas olika rörelser i ansiktsmodellen (VI 7). De olika egenskaperna för nätverket av tyngder och fjädrar motsvarar ansiktsvävnadens egenskaper: olika delar är olika rörliga och musklerna är fästa i ett visst mönster och de kan aktiveras olika mycket. Den stora svårigheten med metoden är att göra en korrekt modell av musklernas egenskaper och av hur de aktiveras utan att den blir för komplex. Risken är att det tar oerhört lång tid att få ansiktet att röra sig. Om man istället förenklar muskelmodellen är det inte längre säkert att det ger en mer korrekt bild av verkliga ansiktsrörelser än den direkta parametriseringen. I de flesta fall där talande ansikten ska användas är den underliggande fysiologin inte heller intressant, utan enbart de resulterande ansiktsrörelserna. Därför är en

allt mer utbredd metod att istället utgå från inspelningar av riktiga personer och låta insamlade data styra hur ansiktet ändras för olika rörelser, så kallad datadriven animering.

Dessa *datadrivna metoder* liknar den direkta parametriseringens sätt att ändra ansiktet utan att modellera ansiktets fysiologi. Skillnaden är att modellens rörelser bestäms genom automatisk analys av insamlade data istället för genom subjektivt bestämda parametrar som man tycker behövs för olika ansiktsuttryck. Med datadriven modellering spelas först ett stort material in med speciella mättekniker. Därefter används någon statistisk metod för att identifiera olika små rörelser, som kan kombineras ihop till de observerade ansiktsrörelserna. Fördelen gentemot den direkta parametriseringen är framför allt att modellen bygger på en verklig talare, så att man med säkerhet vet att rörelserna är naturliga. Detta kan dock också vara en nackdel, eftersom rörelserna från en verklig talare kanske är mindre tydliga än den direkta parametriseringens handgjorda, typiska rörelser. Det är därför viktigt att man väljer talare för den datadrivna metoden så att ansiktsrörelserna blir lämpliga för det tänkta användningsområdet. Om det talande ansiktet exempelvis ska hjälpa hörselskadade med talförståelse, måste talaren ha tydliga munrörelser. En annan nackdel är att mätningarna kräver mycket arbete. Eftersom både mätmetoderna i sig och sättet att analysera den mängd data som man får fram har blir mer och mer effektiva, ökar användandet av datadrivna metoder kraftigt. Idag används mätningar inte bara för att automatiskt bestämma ansiktsuttryck för olika ljud, utan även för olika känslolägen, betoning och skiftande talsituationer. På så sätt får man fram ansikten som blir mer uttrycksfulla och som passar för olika användningsområden.

Vilka mätningar görs för att skapa talande ansikten?

Två typer av information krävs för att göra korrekta modeller av människolika talande ansikten: geometrisk och dynamisk. Med geometriska mätningar får man underlag om ansiktets form för olika visem eller uttryck, likt de nyckelbilder som används vid morfnings- eller interpolationsanimering. De dynamiska mätningarna syftar till att bestämma rörelsen och timingen mellan dessa nyckelpositioner. Till viss del kan man för ansiktet mäta både geometri och dynamik med samma mätningar, medan det för åtminstone tungan krävs att man kombinerar olika tekniker. I det följande exemplifieras detta med de mätmetoder som använts för de talande ansiktena vid KTH. Andra tekniker har använts vid

andra universitet, men de som presenteras här kan antingen anses vara representativa även för andra mätmetoder eller de som i dagsläget ger bäst resultat.

Tungans geometri för olika ljud bestäms bäst genom magnetresonansmätningar. De har två stora fördelar jämfört med röntgenbilder, som förr ofta användes för att mäta tungans rörelser. Den första är att metoden inte utsätter talaren för någon farlig strålning, vilket innebär att det inte finns några etiska begränsningar för mängden mätningar som kan göras. Den andra är att medan röntgenbilden trycker ihop en volym till en platt bild är det med magnetresonans möjligt att ta bilder av enbart en tunn skiva i kroppen (VI 8). Genom att ta bilder av många skivor (VI 9) går det därmed att bygga upp tredimensionella bilder av tungan för olika ljud. Med datadriven statistisk analys skapas därefter en modell (VI 10), vars parametrar kan förändra tungans form från ett ljud till ett annat. Parametrarna kontrollerar exempelvis käkens höjning och sänkning, tungkroppens läge framåt−bakåt och tungspetsens höjning och sänkning. Nackdelarna med magnetresonans för mätningar av tal är att försökspersonen i de allra flesta fall måste ligga ned, att maskinen genererar mycket oljud och att tiden för att ta alla skivor som krävs för att skapa en tredimensionell bild är mycket lång (idag mellan 10 och 40 sekunder, beroende på den bildkvalité som krävs). Det innebär alltså att försökspersonen måste ligga och uttala varje ljud i tiotals sekunder och med mycket bakgrundsbuller. Därmed får man inga upplysningar om rörelser vid tal, enbart om fasta positioner. Det är dessutom inte säkert att personen uttalar ljuden på samma sätt som normalt. Därför krävs även dynamiska mätningar, likt dem som beskrivs nedan. Magnetresonans kan användas även för att mäta ansiktets geometri, men eftersom den är synligt utifrån kan det göras med betydligt enklare metoder.

För att mäta ansiktets rörelser och geometri samtidigt kan man göra inspelningar med två videokameror placerade så att de tar bilder framifrån och från sidan. Eftersom det krävs mycket efterarbetning för att få fram användbar data ur videobilderna är det vanligt att man istället använder ett system som följer markörer. Det har den stora fördelen att man får ut precis den information man önskar direkt om man placerar markörerna på rätt ställen i ansiktet. I det svenska markörföljningssystemet Qualisys klistrar man små reflekterande halva kulor på försökspersonens ansikte (se Figur 3 och VI 11) och filmar därefter med fyra

infraröda kameror när personen talar. Genom att sätta ihop bilderna från de fyra kamerorna ger systemet information om hur varje markör rört sig tredimensionellt (VI 12).

[Figur 3 ungefär här]

Eftersom det hela tiden måste vara fri sikt mellan markörerna och kamerorna kan metoden bara användas för delar som hela tiden är synliga. Därmed kan den inte utnyttjas för exempelvis tungan, som istället måste studeras med specialutvecklad mätutrustning. En spridd metod är elektromagnetisk artikulografi, vilket innebär att små magnetiska spolar fästs på tungan och försökspersonen har en ställning med elektromagnetiska sändare på huvudet (Figur 3 och VI 11). Det magnetiska fältet inuti dessa spolar beror på avståndet till sändarna och man kan därför avgöra hur spolarna rört sig genom att mäta den ström som genereras i spolarna av magnetfältet. På detta sätt kan man få fram hur de delar av tungan där man fäst spolar har rört sig.

Om mätningarna från markörföljningssystemet och artikulografi görs samtidigt, kan man alltså få en mer heltäckande bild av hur ansiktet och tungan rör sig ihop. Inspelningarna kan ge information om ansiktets och tungans rörelser för olika ljudsekvenser (VI 6) eller ansiktsuttryck för olika känslor (VI 13). Modellen för det talade ansiktet får sedan lära sig att återskapa ansiktsrörelserna hos den verkliga talaren. Därefter generaliseras dessa till olika deformationer genom statistiska metoder, så att ansiktsrörelser för nya ljud och uttryck skapas genom att man sätter samman de generaliserade deformationerna.

Även om modellen för det talande ansiktet bygger på inspelningar av en verklig talare innebär det inte automatiskt att ansiktet uppfattas som realistiskt eller naturligt. Det finns därför skäl att avslutningsvis diskutera vad som gör att ett datoranimerat ansikte verkar välgjort eller inte.

Vad krävs för att ett datoranimerat talande ansikte ska vara trovärdigt?

De flesta av de talande ansikten som visats i videoexemplen och figurerna ovan skulle aldrig förväxlas med en verklig människa, möjligen med undantag för någon av de animeringar som bygger på just videoinspelningar av en verklig talare. Att ansiktena är

tydligt animerade beror till viss del på att tekniken ännu inte gör det möjligt att skapa sekvenser som lurar ögat, men kanske till nästan lika stor del på att det finns skäl att inte skapa sådan realism att man förväxlar ansiktet med ett verkligt. Det behöver inte vara ett nödvändigt, och kanske inte ens ett önskvärt, mål att ansiktet ska se fullständigt mänskligt ut. Som tidigare påtalats kan det vara en fördel att det talande ansiktet kan göra överdrivna ansiktsrörelser eller visa sådant som normalt är dolt. Det orealistiska blir i dessa fall bättre än ren realism, om användaren accepterar att det som visas är en virtuell verklighet, där andra regler gäller.

Från datoranimerade Hollywood-filmer är vi vana att allt från fiskar och leksakscowboys till bilar pratar, och åtminstone så länge filmen varar accepterar vi det och bedömer ”realismen” utifrån den animerade världens förutsättningar. Realism innebär i detta fall att de animerade karaktärerna rör sig, artikulerar och har en mimik som överensstämmer med hur vi förväntar oss att deras kroppsrörelser ska vara i den filmens miljö. Karaktärernas ansiktsuttryck upplevs som välgjorda om de både är igenkännbara från människor och har anpassats så att de stämmer överens med den animerade figuren.

För datoranimerade talande ansikten är situationen delvis densamma, delvis annorlunda. Den stora skillnaden ligger i kravet på realism i artikulationen. Medan de animerade biograffilmerna med relativt stor framgång kan dubbas till andra språk, vore det omöjligt för talande ansikten som används för att förbättra talförståelse. Eftersom det ögat ser och det örat hör kombineras, måste informationen i de två kanalerna överensstämma för att inte åhöraren ska bli förvillad. Vid dubbning av animerade filmer säkerställs att röstens och ansiktets känslouttryck motsvarar varandra och dessutom försöker man synkronisera ljud och animation, så att viktiga läpprörelser i animationen och i ljudet överensstämmer. Däremot är det omöjligt att göra en heltäckande matchning mellan ljud och visem, om en sådan över huvud taget fanns i originalanimationen. Ändå är det sällan svårt att förstå de animerade figurerna i filmerna trots att deras läpprörelser inte exakt stämmer med ljudet.

Är det då så viktigt att talande ansikten har helt korrekta ansiktsrörelser i förhållande till ljudet? Svaret är otvetydigt ja. Den stora skillnaden är att vi, när vi tittar på en animerad

filmfigur, inte väntar oss att få hjälp att förstå vad den säger genom att avläsa visem. När vi istället talar med en människa, eller en datoranimerad ställföreträdare, förväntar oss däremot att visemen ska ge tillförlitlig information och därför blir lurade om de inte gör det. Det mest kända exemplet på detta är kanske McGurk-effekten (VI 14), som innebär att om ljudet av ett fonem presenteras tillsammans med bilden av ett annat uppfattas det ofta som ett tredje, mellanliggande. Om ljudet ”ba” spelas upp tillsammans med visemet för ”ga” uppfattas det ofta som ”da”. Orsaken är att bilden visar att det inte kan vara ”ba”, eftersom läpparna inte är sammanpressade, och ljudet att det inte kan vara ”ga”, eftersom frekvenserna inte stämmer. Resultatet av kombinationen blir därför att ett tredje fonem uppfattas, eftersom det någorlunda överensstämmer med båda kanalernas information. Denna effekt, som först upptäcktes av psykologerna McGurk och McDonald 1976, har upprepats många gånger, både med videofilmer av riktiga ansikten och med animerade ansikten. I andra fall är den visuella informationen så tydlig att den dominerar över den akustiska och man uppfattar det fonem som bilden visar, trots att ljudet inte stämmer (VI 14). För videotelefonen Eyephone har man jämfört det talande ansiktets bidrag till förståelsen när animeringen är skapad direkt från taligenkännarens tolkning, som kan innehålla fel, med när tolkningen först korrigerats manuellt. Testet visade att möjligheten för de hörselskadade försökspersonerna att uppfatta orden korrekt var betydligt bättre för den felfria animeringen. Ett huvudkriterium för ett datoranimerat talande ansiktes trovärdighet är därför att akustik och artikulation helt ska överensstämma.

Nästa trovärdighetsfråga gäller videorealism. I de animerade filmer som nämns ovan är det ett mål i sig att karaktärerna ska vara tydligt animerade, likt efterföljare till tecknade filmer, som en kontrast till ”vanliga” filmer med riktiga skådespelare. För datoranimationer i icke-animerade filmer eller i datorspel strävar man däremot efter att göra karaktärer och miljöer så visuellt realistiska som möjligt. Frågan är var någonstans i detta spektrum konstruktörer av datoranimerade ansikten ska sikta. I dagsläget varierar ambitionen mycket mellan olika talande ansikten, från dem som närmast motsvarar tecknade animationer (VI 6), över ansiktsmodeller med en pålagd fotografisk yta av ett ansikte (VI 13) till videobaserad syntes av en mänsklig förlaga (VI 4 och VI 5). I många avseenden är den senaste metoden mest naturtrogen, eftersom den oftast ger en animering som är mest lik en människa som

talar. Det är däremot inte säkert att den därmed är mest trovärdig. Som redan nämnts, får den videobaserade animeringen problem, om inte rätt visem finns i databasen eller om den inspelade personen ändrat huvudställning mellan de olika bilder som satts ihop. När detta händer, uppstår hopp i animeringen. Detta påverkar direkt upplevelsen av ansiktets trovärdighet negativt. Ansiktsmodeller med ett pålagt ansiktsfotografi har inte detta problem, eftersom den underliggande modellen alltid kan skapa kontinuerliga rörelser. Däremot utgör mycket ofta ögon, tänder och tunga ett problem, eftersom det är svårt att automatiskt få dem att se naturliga ut vad gäller skuggning, ytans egenskaper och små ögonrörelser (VI 13). Följden blir därför ofta att dessa ansikten upplevs som obehagliga, på grund av att de inte är enhetliga. Vissa delar uppfattas som naturtrogna, medan andra är avgjort konstgjorda. Framför allt gäller detta om ögonen känns ”döda” och stirrande. Eftersom ansiktsmodeller utan fotografisk yta (som exempelvis i Figur 1) utgör en helhet och aldrig drabbas av diskontinuiteter, kan de trots att de är de minst realistiska ändå vara de mest trovärdiga. De kan till och med vara mer trovärdiga just på grund av att de är mindre realistiska. Förväntningarna på att ett stiliserat ansikte ska röra sig helt naturligt är nämligen mindre än på ett som ser ut att vara en människas riktiga ansikte. Ett andra kriterium för trovärdigheten hos ett talande ansikte är alltså en jämn animationskvalité. Alla ansiktets delar måste se ut att höra ihop i varje bild och de måste röra sig på ett naturligt sätt mellan bilderna.

Nästa möjliga fallgrop för det talande ansiktets trovärdighet är om det stämmer överens med det akustiska talet. Talande ansikten används idag både tillsammans med naturligt tal och med text-till-talsyntes. Naturligt tal kan användas antingen om animeringen sker i realtid, som i Eyephone, eller om yttrandet har spelats in i förväg och animationen sedan anpassas till inspelningen (VI 6). Om naturligt tal används, kan det i vissa fall ge ökad trovärdighet om även ansiktsmodellen är så realistisk som möjligt. Men i så fall krävs att ansiktet överensstämmer med rösten. I exemplet Eyephone skulle det säkerligen upplevas som mycket märkligt om en nära väns röst visas med ett främmande ansikte. De stiliserade ansiktsmodellerna kan däremot accepteras som en neutral representant för alla som ringer, åtminstone så länge talarens och ansiktets kön och ungefärliga ålder överensstämmer. Trovärdigheten kan därmed vara större just därför att ansiktet inte är videorealistiskt.

Om istället text-till-talsyntes används, krävs att naturligheten i ansiktet även finns i ljudet. Det finns idag ett flertal tekniker för att syntetisera tal, som liksom ansiktsanimeringen bygger på en inspelad talare eller på modeller som skapar ljud på konstgjord väg. Dessa olika tekniker har ungefär samma för- och nackdelar som videobaserad eller modellbaserad animering vad gäller naturlighet och möjlighet att variera slutresultatet. Om en talsyntesmetod som låter mindre naturlig används, kan det vara en fördel att använda en mindre realistisk ansiktsmodell. Skälet är att ansiktets naturlighet annars riskerar att medföra att talsyntesen låter än mer onaturlig. Däremot är det mindre problematiskt med den motsatta situationen, med en naturlig röst och ett mindre realistiskt ansikte, eftersom vi är vana vid att mänskliga skådespelare lånar sina röster till animerade karaktärer.

Den information som ges av talet och ansiktet måste också stämma överens, vad gäller exempelvis känslor och attityd. Som redan tidigare nämnts, är kroppsspråk och ansiktsutryck en mycket effektiv signal för vad talaren tycker och känner, ofta mer effektiv än de ord som sägs. Motsägelser mellan ord och ansiktsuttryck förekommer ofta i dialoger människor emellan, när talaren vill dölja sina känslor, men inte riktigt lyckas kontrollera ansiktsuttrycket. I sådana fall tolkar samtalspartnern rimligtvis konflikten mellan orden och ansiktsuttrycket just som att ansiktet avslöjar något mer än det som sägs i orden. Trots att motsägelsen komplicerar tolkningen, är den ur åhörarens synvinkel generellt sett positiv, eftersom den ger ytterligare information.

I kommunikationen med ett talande ansikte är situationen en helt annan, eftersom användaren inte kan förvänta sig att det syntetiska ansiktet ska ha känslor som det försöker dölja. Om exempelvis en animerad lärares beröm ”Vad duktig du är!” åtföljs av ett helt neutralt ansiktsuttryck mister berömmet mycket av sin effekt. Det kan till och med istället tolkas ironiskt. I första hand gäller det alltså att ansiktsuttrycket ska förstärka den information som förmedlas. I andra hand finns även i denna fråga ett samband mellan ansiktets naturlighet i utseende och uttrycksfullheten. Ett verklighetstroget ansikte som inte visar känslouttryck kan lätt uppfattas som mer avståndstagande än ett stiliserat ansikte som inte heller gör det. Det innebär å andra sidan inte att de mer stiliserade ansiktsmodellerna

inte ska visa känslor. Tvärtom – ett viktigt skäl till att karaktärerna i animerade filmer accepteras som levande är just att deras mimik signalerar känslorna mycket tydligt, och ofta överdrivet. Det är möjligt att använda sig av denna övertydlighet för mer stiliserade animerade talande ansikten, men det kan vara mer känsligt att göra det för videorealistiska modeller. Risken finns att det realistiska ansiktets känslouttryck uppfattas som falskt och spelat, eftersom det bedöms efter samma ramar som ett verkligt. Beroende på ansiktsmodellens förmåga att visa känslor som upplevs som menade kan det alltså vara en för- eller nackdel, om det är videorealistiskt.

En något djupare liggande trovärdighetsfråga gäller den bakomliggande kunskap som ansiktet förväntas inneha utifrån utseendet. Det finns ett tydligt samband mellan ansiktets realism och användarens förväntningar på ansiktets kunskaper. Ju naturligare ansiktet ser ut, desto högre blir kraven på att det också ska ha mänsklig förmåga att förstå det användaren säger och ge relevanta upplysningar. Dessutom påverkas användarens eget uppträdande av dessa förväntningar, så att det sätt som han eller hon kommunicerar med ansiktet på beror av hur naturligt ansiktet uppfattas. Med ett realistiskt ansikte kan dialogsystemet få en svårare uppgift, eftersom användaren då uttrycker sig friare och mer komplicerat. Detta motsvarar den effekt som kan observeras i röststyrda upplysningstjänster, där taligenkänningens resultat ofta förvärras när talsyntesen förbättras så att den låter naturligare. Anledningen är att användaren påverkas att prata ”slarvigare”.

Om inte funktionen i systemet kan leva upp till det talande ansiktets naturlighet, blir användaren lätt frustrerad. Beroende på hur avancerad och effektiv tjänsten bakom ansiktet är, kan därför ett mindre naturtroget ansikte vara bättre. Man kan också sänka användarens förväntningar ytterligare genom att istället använda en animerad icke mänsklig karaktär eller till och med ett tecknat gem.

Avslutningsvis bör det påpekas att det möjligen finns en etisk aspekt på hur naturligt realistiska de animerade ansiktena ska göras. Man kan ställa sig frågan om det är eftersträvansvärt att animationen är så trovärdig att den faktiskt kan lura åskådaren att tro att samtalet är med en verklig person. Kanske är det tvärtom önskvärt att det tydligt framgår

att det rör sig om en animation? Svaret är rimligtvis att det beror på användningsområde och på vilken risk det innebär om någon luras. Som jämförelse kan datoranimeringar i spelfilmer ställas emot bildmontage i nyhetsmedia. I det förra fallet snarast förväntar sig åskådaren att animeringen är så verklighetstrogen att man inte kan skilja den från verkliga skådespelare och föremål. I det senare fallet anger man explicit, vilket ofta är övertydligt, att det rör sig om manipulerade bilder. De två fallen skiljer sig i fråga om åskådarens vilja att bli ”lurad” eller inte. Samma problematik kommer säkerligen att uppstå med talande ansikten inom en inte allt för avlägsen framtid. Teknikutvecklingen gör att vi snart möter talande ansikten som, åtminstone utseendemässigt, kan få oss att tro att det rör sig om en levande person. I vissa fall kommer vi antagligen att tycka att det är positivt att det talande ansiktet är så verkligt att det känns lätt att samtala med det. I andra fall kan vi istället uppfatta det som obehagligt att inte veta om det rör sig om en verklig person eller inte.

Vidare läsning

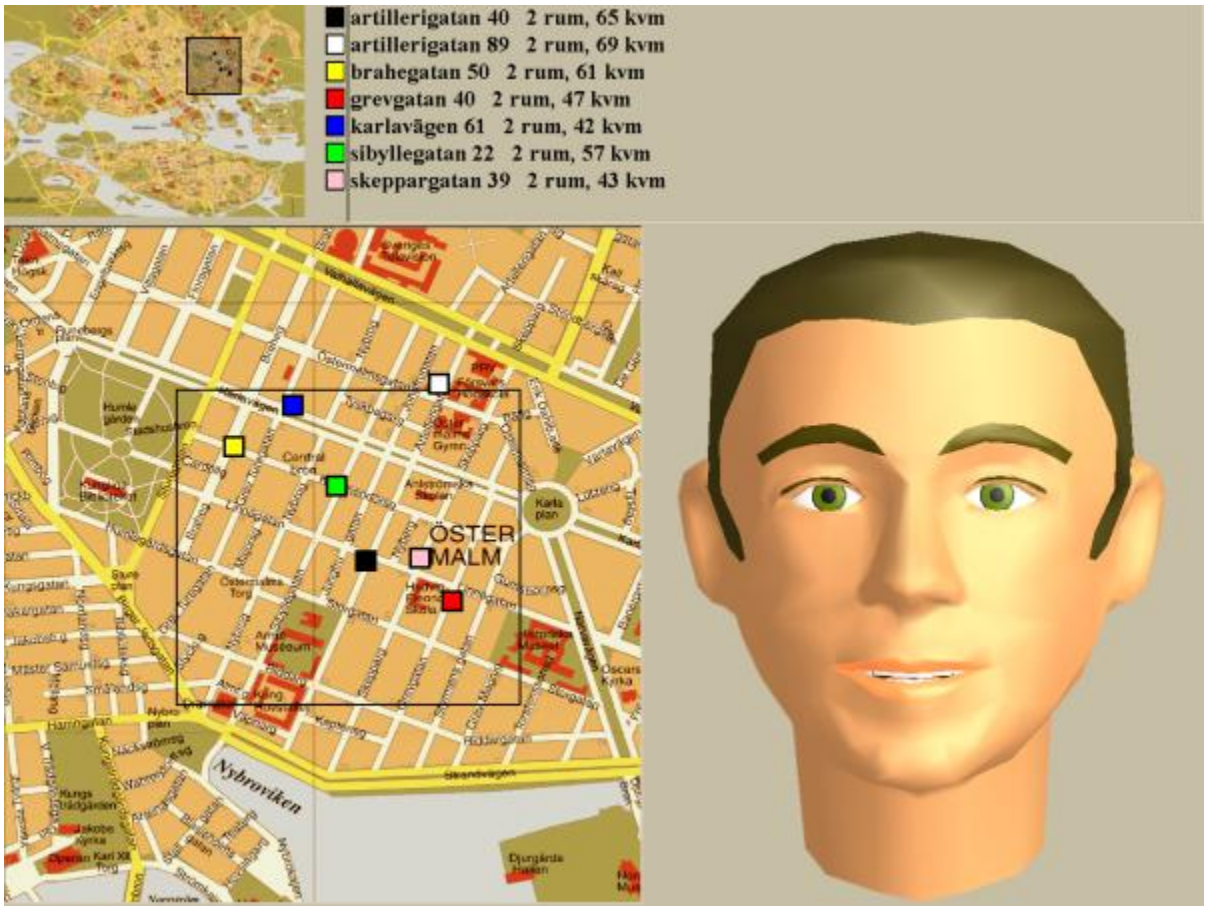
Detta kapitel bygger på akademiska artiklar från i första hand Avdelningen för Tal, musik och hörsel vid Skolan för Datavetenskap och Kommunikation på KTH i Stockholm. Jag har valt att i det ovanstående inte tynga texten med referenser och hänvisar istället den som önskar sådana att ta del av nedanstående tre doktorsavhandlingar, som finns elektroniskt tillgängliga via KTHs bibliotek, http://www.diva-portal.org/kth/theses/. Den första (Beskow) ger en mer djupgående vetenskaplig introduktion till tekniker för att skapa och använda datoranimerade ansikten, den andra (Engwall) beskriver mätningar och modellering av tungan och den tredje (Gustafson) redogör för hur människor interagerar med multimodala dialogsystem.

Beskow, J. (2003). Talking Heads - Models and Applications for Multimodal Speech Synthesis. ISBN: 91-7283-536-2

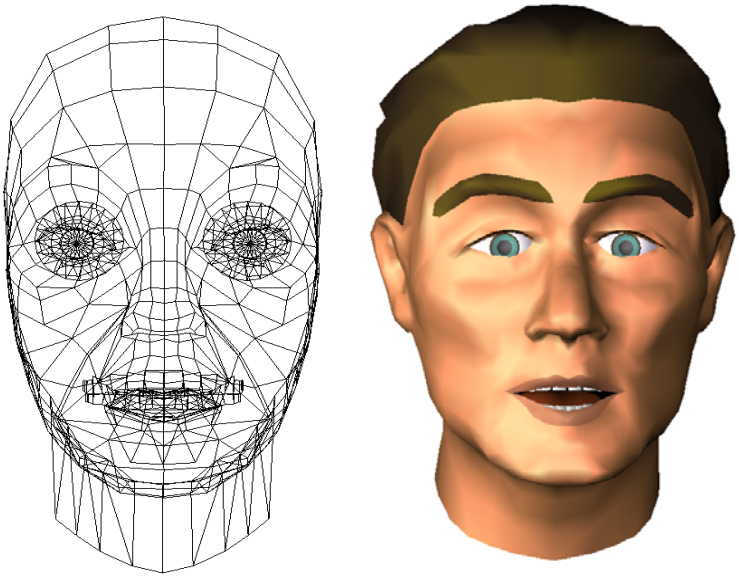
Engwall, O. (2002). Tongue Talking - Studies in Intraoral Speech Synthesis. ISBN: 91-7283-258-9

Gustafson, J. (2002). Developing multimodal spoken dialogue systems. Empirical studies of spoken human-computer interaction. ISSN 1104-5787

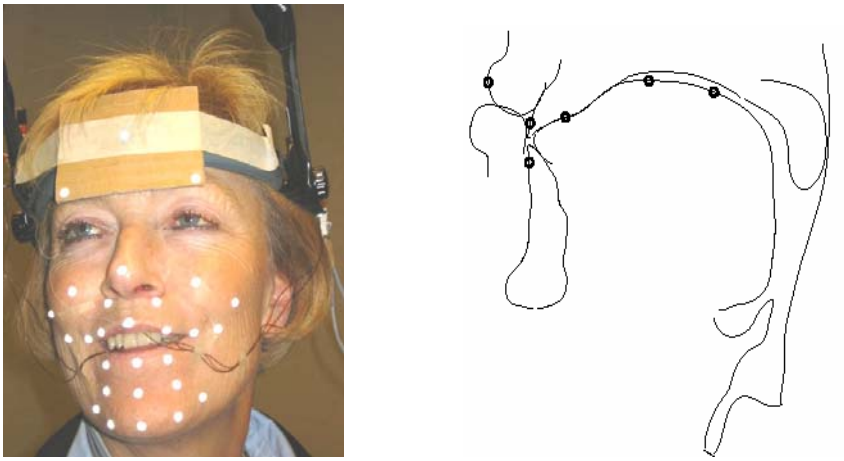
[Figurer. Rangordning: 1) Figur 1; 2) Figur 2 vänster bild; 3) Figur 3 vänster bild; 4) Figur 3 höger bild; 5) Figur 2 höger bild]



Figur 1. Gränssnittet till AdApt-systemet, med översiktskarta, lägenhetsinformation, detaljkarta med lägenhets-ikoner och den virtuelle mäklaren.



Figur 2. Ansiktsmodeller vid KTH. Underliggande rutnät till vänster och med pålagd hudyta till höger.



Figur 3. Utrustning för att mäta ansiktets och tungans rörelser. De vita reflekterande markörerna i vänstra bilden följs i tre dimensioner av infraröda kameror. Huvudställningen håller två elektromagnetiska sändare (utanför bild) som genererar ett magnetiskt fält inuti de sensorer som placeras på tungan och framtänderna. Den högra bilden visar sensorernas placering.