Röstbaserade användargränssnitt

Lars Peterson, lars@flexibel.se
Blekinge Tekniska Högskola
Avdelningen för Interaktion och systemdesign
Innovativa gränssnitt, 7,5 hp

ABSTRACT
Med utgångspunkt i visioner som funnits sedan 50-talet, har jag studerat hur långt utvecklingen nått inom området röstbaserade användargränssnitt. Vad är ett röstgränssnitt och varför ska man använda tal som interaktionsmetod? Hur passar röstgränssnitt in i de trender som idag finns i utvecklingen av innovativa användargränssnitt?

Jag redovisar i vilka användningssituationer röstgränssnitt är lämpliga, gör en jämförelse med grafiska gränssnitt samt beskriver hur röstgränssnitt kan vara en del av multimodala system. Med stöd av aktuell forskning, litteratur och övriga källor, ger jag exempel på användningsområden och applikationer för olika kategorier av röstinteraktion.

Röstgränssnitt bidrar till att vi kan frigöra oss från tangentbord och bildskärm, men vi är fortfarande långt ifrån visionerna om att kunna konversera fritt med en dator om allmänna ämnen. För att nå dit återstår många utmaningar inom bl a lingvistik och kunskap om hur människor för dialoger. I den närmaste framtiden kommer det framförallt att bli i olika typer av expertsystem och nischade produkter och tjänster som vi ser en ökad användning av röst-interaktion, samt som en komponent i multimodala system.

VISIONER
I många science fiction-filmer förekommer det att man pratar med datorerna och får svar i form av en talande röst. T ex datorn ”HAL” i filmen ”2001 – ett rymdäventyr”, datorn ”Mother” i filmen ”Alien”, osv. Detta tycker man ju är fantastiskt – att kunna prata med en dator och få röstsvar tillbaka!

Det som man inte reagerar lika mycker på, iallafall inte jag, är när man i samma eller liknande filmer konverserar med robotar/androider, t ex Ash/Bishop i Alien-filmerna, Mr Data i Star Trek-filmerna, The Terminator i filmen med samma namn, eller C3PO i Star Wars-filmerna... Detta är ju datorer i mänsklig skepnad, och med människor är det inte så konstigt att man för en dialog med, eller hur?

Redan på 50-talet hade man i forskning inom artificiell intelligens ambitionen att utveckla konverserande gränssnitt [23]. Hur långt har man nått idag?

1987 producerade Apple en kortfilm, ”Knowledge Navigator”, som beskrev ett framtidsscenario; En professor förbereder en föreläsning genom att konversera på vanligt språk (”natural language”) med sin personliga agent (”Embodied Conversational Agent”) i datorn [7]. Också Sun Microsystems gjorde en film, ”Starfire”, på liknande tema 1992 [2], där användaren, liksom i ”Knowledge Navigator”, både konverserade med datorn samt interagerade med hjälp av gester. Hur långt ifrån dessa visioner av röstkommunikation med datorn befinner vi oss?

VAD ÄR ETT ANVÄNDARGRÄNSSNITT?
Gränssnitt fanns ju även före datorns intåg. MMI (Man-Machine-Interface) kan definieras som det ”lager” som finns mellan människan och den maskin/verktyg som hon styr och kontrollerar för att utföra en uppgift. Det kan vara en elektrisk maskin med knappar och spakar, men behöver inte vara det. Maskinen kan vara en mekanisk maskin eller ett verktyg som drivs med handkraft (eller annan form av energi), t ex en handpump eller ett roder på en båt, och båtar har ju existerat i tusentals år.

Det krävs alltså två saker av gränssnittet; en möjlighet att påverka maskinen/verktyget genom en ”in-enhet”, samt möjlighet till återkoppling genom en ”ut-enhet”.

Men var går då gränsen för att kalla något ett gränssnitt? Har en yxa ett gränssnitt? Användaren styr eggen genom att hålla i handtaget när han hugger och resultatet syns när träflisorna yr. Men om man går ett par miljoner år tillbaka i tiden, när den moderna människans förfäder började använda verktyg av sten, då hade de första huggverktygen inget påmonterat handtag. Vilket är användargränssnittet då? Finns det ett användargränssnitt?

Om vi återvänder till modern tid och uttrycket HCI (Human-Computer-Interface) så finns det ju flera typer av gränssnitt och in- och utmatningsenheter. I datorernas barndom, för några decennier sedan, påverkade man datorerna genom att ändra i hårdvaran, därefter användes hålkort för inmatningen. Senare kom bildskärm och tangentbord som ju finns kvar än idag, även om användningssättet ändrats en del under resans gång, från ”command-line”-interaktion till det grafiska gränsnittet (GUI) med mus och pekare. Man kan säga att denna utveckling har genomgått fyra olika faser; elektrisk, symbolisk, text- och grafisk fas [8(s.5)].

På senare år har olika typer av sensorer blivit vanligare i gränssnittssammanhang. Sensorer som ser och känner rörelse, sensorer som känner av tryck, beröring och temperaturskillnader. Sensorer som identifierar radio-signaler och ljus. Detta har lett till att vi nu ser nya interaktionssätt växa fram, där man kan styra datorer med

gester och rörelser, i beröring med en yta (t ex multitouch-ytor) eller direkt ”i luften” (t ex ”SixthSense” [37]), eller helt enkelt genom att en sensor känner av en människas närvaro [25(s.149)]. Vi får ett gränssnitt som är mer osynligt än tidigare [8(s.199)], iallafall i fråga om inmatningen, och kan därmed frigöra oss från tangentbord och mus.

En benämning på dessa typer av osynliga gränssnitt är Natural User Interface (NUI) [36], där även Tangible User Interface (TUI) kan räknas in; Att manipulera analoga/fysiska objekt som i sin tur genom olika sensorer fungerar som inmatningsenheter till en dator är en fascinerande tanke. ”Tangible Interaction” är verkligen direktmanipulation!

Vad är då gränssnittet i dessa fall, när tangentbord och mus inte längre är aktuella? Jag anser att det är de ”osynliga” sensorerna som är det egentliga inmatningsgränssnittet, inte själva objekten, gesterna eller rörelserna i sig. Det är därför man inte kan tala om användargränssnitt om man manipulerar vanliga analoga/fysiska objekt som INTE är kopplade till en maskin/dator, eftersom syftet då inte är att styra en maskin. Där finns inget ”Tangible User Interface”.

I vanliga skrivbordsdatorer och bärbara datorer finns ofta, förutom tangentbord, skärm och mus, även mikrofon och högtalare som gränssnitt, även om dessa inte används så mycket ”till vardags” för in- och utmatning av data. Det mest vanliga är kanske att ett varningsljud hörs från högtalarna för att påkalla uppmärksamhet då användaren gjort något som inte fungerade som det var tänkt. Mikrofonen används till direktkommunikation med andra människor, t ex via Skype, som alternativ till vanlig telefoni, samt högtalarna för att lyssna på musik, eller lyssna på ljudet från en film som spelas upp på skärmen, som alternativ till att lyssna i en musikspelare respektive att se film i en dedicerad videouppspelningsapparat. Datorn konvergerar, som Dourish [8(s.194)] påpekar, till en multi-medial maskin med internet som kommunikationskanal.

Att använda ljud för in- och utmatning av data tillsammans med det grafiska gränssnittet är exempel på ett multimodalt gränssnitt, att man använder flera olika in- och utmatnings-sätt samtidigt eller alternerande. Dock är nästan alltid bildskärmen närvarande som utmatningsenhet, och i många fall används tal endast för ge kommando på exakt samma sätt, och med samma vokabulär, som när man ger kommando via tangentbord och mus.

RÖSTGRÄNSSNITT

I ett renodlat röstbaserat användargränssnitt består både inmatning och återkoppling av enbart talat språk (”natural language”), dvs t ex engelska eller svenska, i motsats till ett konstruerat språk som t ex ett programmeringsspråk. Sedan förkommer det naturligtvis system eller applikationer där enbart inmatningen alternativt enbart återkopplingen består av talat språk.

Gränssnittet i interaktiva system där användaren och systemet konverserar via talat språk, benäms på engelska Voice User Interface (VUI) [11(s.12)]. Det finns många synonymer till detta uttryck, vilka ofta inte ens innehåller ordet ”interface”, t ex Spoken Language Systems, Automatic Telephone Dialogues eller Spoken Dialogue Systems [11(s.4)]. Namnen avslöjar att många applikationer med röstgränssnitt har sitt ursprung i telefonbaserade applikationer, där ordet gränssnitt inte varit ett dominerande uttryck, eftersom man ofta inte betraktat konstruktionen av röstgränssnitt som design av ett användargränssnitt på motsvarande sätt som designers gjort med grafiska gränssnitt.

Det behövs kompetens inom många olika språkforsknings-områden för att konstruera dugliga konverserande interaktiva system – lingvistik, filosofi, sociologi och psykologi [11(s.3)]. För att designa ett ”Spoken Dialogue System” krävs flera avancerade systemkomponenter, bl a taligenkänning, språkförståelse, dialoghantering, databas-kommunikation, responsgenerering och text-till-tal syntes [23(s.91)].

VARFÖR ANVÄNDA TAL SOM INTERAKTIONSMETOD?

Vid vilket tidpunkt i historien språket utvecklades råder det delade meningar om. Det finns egentligen bara spekulationer, men en teori är att språket uppstod hos Homo sapiens för hundratusentals år sedan. Bland ytterligheterna förekommer teorier att språket utvecklats för miljoner år sedan hos Homo habilis, eller att det först dök upp hos den moderna människan för kanske 50.000 år sedan [35]. I vilket fall som helst är det talade språket, tillsammans med med gester och kroppsspråk, det mest naturliga sättet att kommunicera människor emellan. Genom att även i kommunikationen med datorer använda dessa för människan naturliga sätt att kommunicera uppnår vi det som Dourish [8(s.102)] benämner ”embodied interaction”.

Om man ska jämföra ”naturligt språk” med gester som interaktionsmetod, har gesterna den nackdelen att det krävs att användaren lär sig vilka gester som kan användas, och hur de ska utföras, för att uppnå det han/hon önskar. I ett system som kan tolka ”naturligt språk” behöver användaren inte lära sig ett nytt språk, varken verbalt eller ett språk av gester.

ANVÄNDNINGSSITUATIONER

En anledning till att interagera med datorer via röstgränssnitt, är att det frigör oss från tangentbord och bildskärm. Telefoner finns överallt men inte bildskärmar – inte än iallafall. Även om bildskärmarna i mobiltelefoner blir bättre och större, finns det situationer där våra händer eller ögon är upptagna med annat. Vi kan då via röstgränssnittet kommunicera med expertsystem i situationer där man t ex använder verktyg eller av olika anledningar måste bära handskar eller vantar, eller när man framför ett fordon.

Som hjälpmedel för handikappade, som inte kan nyttja sina händer eller ögon, är röstgränssnittet ett alternativ för att interagera med olika system.

RÖSTGRÄNSSNITT KONTRA GRAFISKT GRÄNSSNITT

Naturligtvis finns det nackdelar med röstgränssnitt, men man måste komma ihåg att ett röstgränssnitt inte är ett ”audio-GUI”, dvs det går inte att direkt konvertera ett grafiskt gränssnitt till ett genuint röstgränssnitt, och vice versa [11(s.211)]. Gränssnitten kompletterar varandra, och i vissa situationer är det en fördel att använda båda sätten att interagera med ett system, genom ett multimodalt gränssnitt.

Den förmedlade informationen i ett röstgränssnitt är temporär och linjär; när du tagit del av informationen från datorn finns den bara kvar i ditt minne (förhoppningsvis), och du måste be datorn upprepa informationen om du vill få tillgång till den igen. Information i form av text och bilder på en bildskärm visas på skärmen så länge du önskar.

Likaså kan du på bildskärmen snabbt få en överblick och kan med blicken sortera mycket information samtidigt, samt också jämföra data. I ett röstgränssnitt kan du bara lyssna på ett meddelande i taget – vi klarar inte av att lyssna på flera röster samtidigt.

Röstgränssnitt är inte lämpligt att använda i alla situationer. På allmänna platser eller i andra situationer där omgivningsljud kan vara störande, kan det vara svårt för ett system att uppfatta vårt tal, och även för oss att uppfatta datorns svar. Om vi inte vill att andra ska höra vår dialog med datorn, är publika platser inte heller den bästa platsen att nyttja röstinteraktion.

En nackdel med renodlade röstgränssnitt, t ex i enklare telefonapplikationer med hierakiska menysystem, eller i röstkontrollerade system där enbart inmatning sker genom att använda rösten och det heller inte finns någon visuell återkoppling, är att det kan vara svårt att veta i vilket tillstånd systemet befinner sig i, eller vilka kommandon som är tillgängliga. I mer avancerade konverserande system är detta inget större problem, eftersom systemet här är mottagligt för en dialog som inte följer en exakt fastlagd tågordning.

I en mer avancerad röstbaserad applikation kan man också direkt få tillgång till önskad information – man slipper att stega igenom flera menyer och sidor som man kanske är tvungen till i ett grafiskt gränssnitt.

För enklare röstkontrollerade system, vilka finns i t ex hem eller bilar, som ständigt måste lyssna efter kommando, kan det vara ett problem för systemet att avgöra om talet från användaren är ett kommando eller han/hon pratar med någon annan människa. Ett sätt att lösa detta är att tilltala datorn med ett namn, som i Star Trek, där man adresserar datorn med ordet ”Computer”, varvid datorn ”vet” att det som sägs efter det är ämnat för denna.

MULTIMODALA ANVÄNDARGRÄNSSNITT

I ett multimodalt system kan man kombinera det bästa från olika typer av gränssnitt, och använda de sätt att interagera som är mest lämpliga för uppgiften man ska utföra och som passar bäst i den kontext man befinner sig i. Några enkla exempel är att man på en bildskärm kan peka på ett objekt och sedan säga ”berätta för mig om detta objekt” [26(s.163)], eller peka på ett objekt och säga ”röd” [24]. Ett annat exempel är att kunna peka på en plats på en karta på bildskärmen och fråga efter navigationsinstruktioner [19]. Ytterligare ett användningsområde kan vara att man ger ett röstkommando för att direkt komma till en sida där man sedan bearbetar något via det grafiska gränssnittet.

I ett multimodalt gränssnitt i t ex en ”smartphone”, där inmatning både via det grafiska gränssnittet och via tal är möjligt parallellt, kan det vara förvirrande och omständligt att använda tal på det grafiska gränssnittets premisser – risken finns att röstgränssnittet endast blir ett ”audio-GUI” med exakt samma kommandon fast istället uttalade med rösten. Likaså kan det vara önskvärt att i miljöer med mycket omgivningsljud kunna stänga av röstgränssnittet. En lösning är att man gör två olika grafiska gränssnitt som återkopplar och indikerar ”affordance” – ett gränssnitt anpassat för inmatning via pekskärmen och ett anpassat för röstinmatning, det senare befriat från knappar, rullister, ”drop-down”-menyer och andra typiska GUI-objekt [19]. De två olika inmatningssätten kan då inte nyttjas parallellt, men tanken är att det på ett mycket enkelt och snabbt sätt ska gå att växla till det andra inmatningssättet.

ANVÄNDNINGSOMRÅDEN OCH APPLIKATIONER

Röstgränssnitt används inom en mängd olika områden. Applikationer där enbart röstinmatning används, applikationer där tal endast nyttjas som återkoppling, andra kombinationer av text och tal i in- respektive utmatning, samt i system där all interaktion sker enbart med tal. Från enkel röststyrning av apparater i hemmet och funktioner i bilen, eller att mobiltelefonen på röstkommando ringer upp en person som finns i telefonboken, till att t ex en reparatör/tekniker för en dialog med en ”conversational agent” för att få guidning vid avgörande arbetsmoment.

Telefonbaserade system dominerar idag bland system med röstinteraktion. Av praktiska skäl används på webbsidor ofta textinmatning till konverserande agenter, men att därifrån istället använda tal är inte ett stort steg. Nielsen [24] beskriver det så här; ”All that voice does is let users speak, rather than write, commands and parameters. A small part of the puzzle, indeed.”

Röstkommunikation används inom kundservice, help-desk, vid guidad försäljning och teknisk support [20]. Om du går in på t ex IKEAs brittiska webbsajt, hjälper en blond ”Anna” dig att navigera på sajten och att hitta rätt produkter genom att svara med talad engelska [15]. Konverserande agenter kan också användas i utbildningssituationer, t ex som rådgivare vid problemlösning [9].

Andra av oss här i Sverige kända telefonapplikationer är t ex biljettbokning hos SJ (ring telefonnr 0771-757575 och välj ”självbetjäning”) samt Telias kundtjänst (telefonnr 90200), vilka kan kategoriseras som en enklare form av konversation. Ofta är det dessa typer av sekvensiella system med hierarkiska menyer, som allmänheten känner till och tänker på när man pratar om röstbaserad interaktion [11(s.15)]. System som egentligen bara är snäppet bättre än de som styrs av tryckningar på telefonens nummerknappar och som får folk att bli frustrerade [16].

Varför då överhuvudtaget använda telefonbaserade system? Är det inte bättre att istället direkt få prata med en levande människa? För företagen som erbjuder tjänsterna handlar det naturligtvis om ekonomi och effektivisering. Men det finns också en annan aspekt; Att ha tillgång till olika expertsystem och guidningar alltid och i alla situationer kan också effektivisera och underlätta den informationssökandes tillvaro.

EXPERTSYSTEM

En trend, både enligt Dourish [8(s.195)] och Nielsen [24] är att vi i framtiden får fler små specialiserade interaktiva digitala produkter i vår vardag – ”information appliances”. Produkter som har avgränsade användningsområden till skillnad från skrivbordsdatorn som vi använder till det mesta. Sådana specialiserade produkter kan vara ytterst lämpliga att interagera med via rösten, menar Nielsen [24].

Det finns inga röstsystem som man kan hålla en konversation med om allmänna ämnen [20], eftersom sådana system är alltför komplicerade att konstruera. Detta är en anledning till att dialogbaserade system är specialiserade. De är expertsystem som kan hantera frågor inom sitt område. Men egentligen finns det ingen större anledning att sträva efter något annat. I de flesta situationer är vi ute efter information eller hjälp inom ett visst avgränsat område, och inom detta område används vanligtvis en viss vokabulär och vissa frågor förväntas avhandlas. Detta underlättar vid konstruktionen av det dialogbaserade systemet.

OLIKA KATEGORIER AV RÖSTINTERAKTION

”Icke-tal”

Den enklaste formen av röstinteraktion är egentligen där ord inte uttalas, utan man använder rösten till att frambringa andra ljud. Exempel på områden med sådana applikationer, där man använder olika tonlägen och styrka på ljudet som man frambringar med rösten, är t ex inom spel, konst eller assisterande teknologi [18]. Ett annat exempel är att kunna styra en hushålls-mixer genom att skrika åt den [25(s.148)].

En applikation, som egentligen mer är en övervakningsapplikation än en frivillig människa-dator-interaktion, är följande:

För att snabbt försöka upptäcka personer med symptom på smittsam influensa, har ett belgiskt företag utvecklat en hostdetektor, som enligt företaget kan höra skillnad på vanlig hosta och en sjuklig [32]. Detektorerna kan antingen placeras ut i mängder på en flygplats, där en misstänkt smittbärare snabbt kan identifieras, eller så kan detektorerna byggas in i mobiltelefoner, som då kan varna om antingen mobilprataren eller någon i omgivningen hostar på ett misstänkt sjukligt sätt.

Kommandon

Den enklaste formen av röstinteraktion där tal används, är när man ger korta kommandon till system som återkopplar genom att bekräfta och utföra kommandot. Röstkontroll av grafiska gränssnitt i datorer, t ex VoiceOver eller MacSpeech för MacOS samt Dragon NaturallySpeaking för Windows, eller röststyrd uppringning i mobiltelefoner, är några exempel. Att skicka sms genom att diktera meddelandet och därefter ge kommandot att skicka det är också möjligt [13].

Enklare röststyrda telefonapplikationer kan också räknas till denna kategori. Sådana applikationer är väldigt lika de system som man styr med tryckningar på telefonens knappar. De är uppbyggda enligt samma hierarkiska struktur, med ett tidskrävande en-fråga-ett-svar-dialog. Det kan t ex utspela sig som följer [11(s.15)]:

Systemet: För information om transport, säg ”transport”. För information om underhållning, säg ”underhållning”. För information om väder, säg ”väder”.

Uppringaren: Transport.

Systemet: För information om flyg, säg ”flyg”. För information om bussar, säg ”bussar”. För information om tåg, säg ”tåg”.

Uppringaren: Flyg.

Systemet: För information om ankomster, säg ”ankomster”. För information om avgångar, säg ”avgångar”.

osv.

En bilinteriör är en tacksam miljö att implementera röststyrning i, därför att föraren alltid sitter på samma ställe och mikrofon och högtalare är därmed lättplacerade. Mercedes-Benz har utvecklat röstsystem för styrning av luftkonditionering, navigationsutrustning, musikanläggning mm [12].

”SmartWeb” heter en produkt som utvecklats tillsammans med Mercedes-Benz bilar och BMW motorcyklar [33]. ”SmartWeb” är en multimodal styrd webbservice där användaren kan få guidning och hjälp inom många olika områden.

Ford har i samarbete med Microsoft utvecklat applikationen ”Ford SYNC” som ger möjlighet att röstkontrollera mobiltelefoner och musikspelare från förarmiljön [34]. Applikationen finns implementerad i ett flertal Fordmodeller.

Exempel på röststyrda apparater i hemmet är t ex bild- och ljudanläggning, luftkonditionering, värmeanläggning, lampor, väckarklocka och larm [30].

Man kan med anpassad programvara även styra apparaterna på distans genom att ringa upp dem från en telefon när man inte är hemma. Enligt en tillverkare, som döpt sin applikation till HAL2000, är det möjligt att ge mer komplexa kommandon än enstaka ord; ”Every Saturday and Sunday Turn Living Room Lights On at 6 pm for Three Hours” [29]. I detta fall återkopplar systemet med att säga att det registrerat den önskade åtgärden.

Uppläsning

Alla känner nog till telefonapplikationer där man styr informationsflödet med att trycka på telefonens knappar. ”Tryck 1 för att komma till kundtjänst, Tryck 2 för support”, osv. Därefter kommer en ny nivå som presenteras på samma sätt. Dessa typer av applikationer är ju begränsade på så sätt att man bara har tolv knappar att välja på, och det kan vara väldigt tröttande för användaren att nå dit man vill om målet ligger långt ner i hierarkin. Ibland kan det hända att man som uppringare inte tycker att något alternativ passar in på det ärende man har. Vad gör man då?

I slutet av 90-talet trodde många företag inom röstteknologi att det inom några år inte skulle vara nödvändigt med en grafisk webbläsare i mobiltelefonen, utan att man istället helt enkelt skulle få tillgång till all information på webben via röstinteraktion. Denna entusiasm dog ut när IT-bubblan sprack. Vem vill lyssna på en dator som återger en webbsida via telefonen? Det är nästan jämförbart med att se en videofilm återgiven i en tidskrift, med en bildruta per sida [11(s.209)]. Vem vill se en film på det viset? Företagen gjorde misstaget att inte betrakta röstgränssnitt som en egen unik typ av användargränssnitt, utan som ett ”audio-GUI”.

Däremot finns det företag som fortfarande tror på att via röstgränssnitt komma åt den data som finns BAKOM det grafiska gränsnittet, dvs själva informationen, inte den grafiska yta som informationen presenteras på via webbläsaren. IBM har en vision att det inom fem år ska gå att surfa på Internet genom att bara använda rösten [14]. Man kallar tekniken för ”Spoken Web”. Man tror att tekniken främst kommer att ha framgång i de länder där datorer inte är vanligt förekommande. Det är bara 17% av världens befolkning som idag har tillgång till Internet. Tekniken innebär att man med en vanlig telefon både ska kunna lägga upp och ta del av information på ”VoiceSites”. Detta betyder att även personer som inte kan läsa och skriva ska kunna använda Internet genom en parallellt WWW, som IBM har döpt till The World Wide Telecom Web (WWTW) [1, 17].

Konversation

”Spoken Dialogue Systems” kan ses som avancerade applikationer där gränsnittet mellan människan och datorn tillåter en dialog på talat språk i form av ett för människan någorlunda ”naturligt” språk. De enklaste systemen är uppbyggda enligt en-fråga-ett-svar-principen. I de mest avancerade går det att föra en betydligt mer utförlig dialog. De enklare systemet kan kanske bara ”förstå” några få ord, som ”ja” och ”nej”, medan de avancerade kan tolka långa meningar [23(s.92)].

Att kunna föra en dialog innebär en stor fördel. Informationen blir snabbare tillgänglig när användaren slipper gå igenom långa sekvenser av frågor och svar. Det är också viktigt att noviser ska kunna använda avancerade röstsystem, utan att först behöva lära sig en viss vokabulär. Detta blir möjligt tack vare dialogen, som ger stöd i hur användaren ska få tillgång till informationen. I system som kan kategoriseras mittemellan en-fråga-ett-svar-principen och dialoger med fullödigt naturligt språk, kan man låta användaren först lära sig en begränsad vokabulär av styrord för att enklare kunna interagera med systemet på ett effektivt sätt [27].

Ett sätt att strukturera konverserande system är att dela in dem i följande tre olika nivåer [23(s.92)]:

(1) ”finite state”-baserade system.

(2) ”frame”-baserade system.

(3) ”agent”-baserade system.

”Finite state”-baserade system är den form av system som hittills varit mest frekvent i kommersiella applikationer:

System: What is your destination?

User: London.

System: Was that London?

User: Yes.

System: What day do you want to travel?

User: Friday.

System: Was that Sunday?

User: No.

System: What day do you want to travel?

[23(s.93)].

Systemet styr dialogen som följer en förutbestämd sekvens genom ett antal alternativa vägar. Det krävs bekräftelse efter varje lämnad uppgift, och inmatningen består av enstaka ord. Fördelen med detta system är att man i förväg kan definiera vilka ord systemet behöver känna igen. Nackdelarna är att dialogen är begränsad och tidsödande. Användaren kan inte ta initativet.

I ”frame”-baserade system sparas uppgifterna från användaren i en mall (”template”) som fylls i efter hand som systemet uppfattar informationen:

System: What is your destination?

User: London on Friday around 10 in the morning.

System: I have the following connection…

[23(s.93)].

Dialogen måste här inte följa en förutbestämd väg och systemet ställer frågor endast om det saknar uppgifter för att leverera önskad information. Bekräftelsen kan dröja tills systemet har uppfattat all information.

Följande exempel på ett ”frame”-baserat system är från ”Philips Automatic Train Timetable Information System” [23(s.98)], som tillhandahåller information om tåg-förbindelser mellan 1.200 tyska städer (översatt till engelska):

System: Good morning. This is the automatic timetable information service. How can I help you?

User: I need a connection from Hamburg to Munich.

System: When would you like to go from Hamburg to Munich?

User: Next Friday at eight o’clock.

System: So you want to travel on Sunday the 20th of November at eight o’clock a.m.?

User: No, on Friday.

System: So you want to travel on Friday the 18th of November at eighto’clock a.m.?

User: Yes.

System: There is the following connection: With IntercityExpress 793, departure from Hamburg at 8:02, arrival in Munich at 14:05. Do you want me to repeat the connection?

User: No, but I need a train for the return trip on Sunday.

System: At what time would you like to go from Munich to Hamburg on Sunday the 20th of November?

User: I have to arrive in Hamburg late in the evening.

System: So you want to arrive in Hamburg between nine p.m. and eleven p.m.?

User: Yes.

System: There are three connections. This is the first one: With IntercityExpress 880, departure from Munich at 15:54, arrival in Hamburg at 21:58. Do you want the next connection?

User: No, thanks, that one is fine.

System: Do you need yet another connection?

User: No, thank you.

System: We wish you a pleasant journey. Thank you for calling. Good bye.

I ”Agent”-baserade system kan kommunikationen vara än mer komplex och initiativet till nya vägar i dialogen kan växla mellan systemet och användaren på ett mer naturligt sätt. Här är ett exempel [23(s.94)]:

User: I’m looking for a job in the Calais area. Are there any servers?

System: No, there aren’t any employment servers for Calais. However, there is an employment server for Pasde-Calais and an employment server for Lille. Are you interested in one of these?

Ett annat exempel på ”Agent”-baserade system är taget från ”The Circuit-Fix-It Shop System” [23(s.100)] och beskriver hur system och användare samarbetar för att lösa ett problem med en elektronisk krets. Information från båda håll krävs för att lösa problemet. Agenten sitter inte inne med all kunskap utan måste få hjälp från användaren:

System: Is there a wire between connector 84 and connector 99?

User: It is not there.

System: Add a wire between connector 84 and connector 99.

User: I need help.

System: Obtain a wire.

User: Done.

System: Connect metal of one free end of the wire to connector 84.

User: How?

System: Bend back connector 84.

User: Done.

System: Insert free end of wire into connector 84.

User: Wire between 84 and 99 is connecting.

System: What is the LED displaying?

EMBODIED CONVERSATIONAL AGENTS

Vid konstruktion av ”Conversational agents” är det viktigt att tänka på hur agentens talade röst framstår. Det kan ha betydelse för framtoningen och varumärket för det företag eller organisation som agenten representerar. Bara genom att lyssna på en röst, utan att man ser den som talar, kan man tolka in många saker. Man avläser personlighet genom framtoning, tonläge och vokabulär. Konstruktören måste också bestämma röstens kön, och om den ska vara syntetisk eller verklig.

Uttrycket ”Conversational agent” har synonymer som t ex ”Artificial conversational entity” och ”Chatbot”/ ”Chatterbot”. En ”Embodied conversational agent” (ECA) är en ”Conversational agent” som representeras av en grafisk framställning på en bildskärm. Det kan vara en 2D- eller 3D-representation i form av en människa, djur eller fantasiväsen. Här tillkommer alltså ytterligare en faktor att

ta ställning till för konstruktören. Undersökningar har visat att människor föredrar en grafisk representation som motsvarar den nivå på uppgift som agenten ska bistå att lösa, och att representationen stämmer in i sammanhanget [10].

Syftet med den grafiska representationen är att göra interaktionen mer engagerande och agenten mer trovärdig [3(s.651)], bl a genom att återge uttryck i ansikte och kroppsspråk. Forskning har visat att människor ändå behandlar datorer som sociala varelser, även utan en grafisk representation [10(s.210)]. Undersökningar i utbildningssammanhang har dock visat att närvaron av en ECA inte säkert har någon effekt på inlärningen eller uppgiften som ska utföras. Det har bara uppmuntrat att använda ECA:n som hjälpfunktion [28]. Andra personer, bl a Ben Shneiderman, menar att man inte alls ska bry sig om att försöka representera en applikation med en grafisk framställning, utan man ska behandla datorer som datorer, inte som människor [11(s.305)].

Representationen kan också bestå av ett fysiskt objekt, men om man inte beger sig in på robotområdet, missar man återkopplingen i form av kroppsspråk och ansiktsuttryck. Ett exempel på en fysisk framställning av en ECA är en livsstilsrådgivare representerad av en Nabaztag, en trådlös plastkanin [31].

REA (Real-Estate Agent) är namnet på en ECA utvecklad på MIT [5]. Här försöker man gå ett steg längre. REA representeras av en hel människoliknande kropp återgiven i en 3D-miljö på en stor skärm. Användaren kan föra en dialog med REA om mäklartjänster, där objekten man samtalar om, som hämtas från en databas med lediga bostäder i Boston, kan visas på skärmen bakom REA. REA kan känna av närvaro och användarens kroppspositioner och handrörelser. REA kan använda ett relativt komplext kroppsspråk, som positioner, gester och ögonrörelser, och hon kan söka ögonkontakt. I dialogen med användaren kan initiativet växla, ordet överlämnas på ett ”naturligt” sätt, och hennes tonläge i rösten kan skifta beroende på situation.

”The Loebner Prize” [21] och ”The Chatterbox Challenge” [6] är två tävlingar där man premierar de bästa ”Conversational agents” enligt vissa kriterier. På respektive webbsidor finns länkar till diverse ”Conversational agents”.

SLUTSATS

Det har funnits en optimism kring att utveckla användarbara ”Spoken Dialogue Systems” sedan 50-talet. Att det har tagit längre tid än man trodde, beror på att man varit alltför teknikfokuserad, när det snarare är inom lingvistik och kunskap om hur människor för dialoger som knäckfrågorna finns. Datorn ska inte bara uppfatta ord, utan också förstå användarens mål. Det hjälper inte att man bygger superdatorer. 1997 vann superdatorn ”Deep Blue” från IBM över schackvärldsmästaren Kasparov, och nu drygt ett decennium senare satsar IBM på att deras superdatorteknik ”Watson” ska kunna spela Jeopardy [22]. Men att förstå och svara på frågor, i princip fungera som ett gigantiskt uppslagsverk, är inte det samma som att kunna föra en fruktbar dialog.

Kapaciteten i alla ingående komponenter för röstsystem – taligenkänning, språkförståelse, dialoghantering, databaskommunikation, responsgenerering och text-till-tal syntes – kommer så sakta att förbättras, vilket leder till att vi får se fler röstbaserade system med bättre noggrannhet och ett bättre flöde i dialogen [20(s.15)]. För att nå framgång är det viktiga hur alla dessa komponenter integreras tillsammans. "Conversational agents” som idag använder sig av text för in- eller utmatning, kommer i framtiden istället att hantera tal på samma lingvistiska nivå som de idag hanterar text [20(s.16)]. Att tala till datorn istället för att skriva in text, är ett litet problem i sammanhanget.

Designers måste komma ihåg att röstgränssitt inte bara är ”audio-GUI”. Men eftersom även gestbaserade interaktionsmetoder nu utvecklas, lär vi oss att komma bort från låsningen att designa för GUI. Då lär sig designers också att betrakta röstgränssnitt som ett eget unikt gränssnitt. I likhet med att gestbaserade interaktionsmetoder kommer att bli vanligare, kommer även röstbaserad interaktion öka, vilken bidrar till att vi får ”osynliga gränssnitt”. Multimodala gränsnitt, med rösten som en interaktionsmetod, kommer att dyka upp i många sammanhang.

En anledning till att ständigt förbättra och förenkla vår vardag, där röstgränssnitt är en sätt, är som Hugh Dubberly (en av de personer som gjorde ”Knowledge Navigator”) säger – ”det viktiga är hur tekniken kan nyttjas i våra sociala liv, vad tekniken kan göra med och för människor”. Stephen Intille på MIT anser att utvecklare bör ställa sig frågan ”Kommer applikationen att förbättra människors livskvalitet?” [4].

Men det kommer att dröja innan visionerna i ”Knowledge Navigator” och i sci-fi-filmer är infriade, att kunna föra en dialog med en ECA om vitt skilda ämnen. Istället blir det inom expertsystem (guidning, support) och nischade produker och tjänster där ”Spoken Dialogue Systems” kommer att ha en funktion för effektivt samarbete mellan människa och maskin.

Utmaningarna inför framtiden är många och olika, här är några exempel;

Hur löser vi problem med omgivningsljud och hur ska vi adressera datorn i sådana miljöer? Hur får vi systemet att känna av användarens tonläge och humör? Hur ska systemet kunna utläsa användarens kunskapsnivå och önskemål? Och som Jokinen [16] frågar sig; Hur bygger vi in etiska överväganden i agenternas beslut?

Jag har tidigare varit skeptisk och ofta blivit irriterad på de få applikationer där jag stött på röstinteraktion. Men nu när jag vet som är möjligt, är min nyfikenhet väckt och det ska bli spännande att se vad som sker inom röstinteraktion i framtiden. Och vem vet hur framtiden kommer att se ut?

Även om tidpunkten för när filmen ”2001” har passerats, så utspelar sig både Star Trek och Star Wars i en tid som ligger framför oss...

Lars Peterson, 2009-06-01.

REFERENSER

1. AGARWAL, S. K., KUMAR, A., RAJPUT, N., NANAVATI, A. A., RAJPUT, R., 2008. *The World Wide Telecom Web Browser*. New York, NY: ACM. International World Wide Web Conference; 1121–1122.
2. ASK TOG. Interaction Design Solutions for the Real World. Nielsen Norman Group. Starfire. *A Vision of Future Computing,* Sun Microsystems, 1992. http://www.asktog.com/starfire.
3. BENYON, D., TURNER, P., TURNER, S., 2005. *Designing interactive systems : people, activities, contexts, technologies*. Harlow, England : Pearson Education.
4. BERGMAN, E., LUND, A., DUBBERLY, H., TOGNAZZINI, B., INTILLE, S., 2004. *Video Visions of the Future: A Critical Review*. New York, NY: ACM. Conference on Human Factors in Computing Systems; 1584–1585.
5. CASSELL, J., 2000. *Embodied conversational interface agents*. New York, NY: ACM. Communications of the ACM; Volume 43, Issue 4, 70–78.
6. CHATTERBOX CHALLENGE (CBC). *The Chatterbox Challenge 09*. http://www.chatterboxchallenge.com.
7. DIGIBARN Computer Museum. *The Knowledge Navigator concept piece by Apple Computer*, 1987. http://www.digibarn.com/collections/movies/knowledge-navigator.html.
8. DOURISH, P., 2004. *Where the action is : the foundations of embodied interaction*. Cambridge, MA : MIT Press.
9. ELZWARE Conversational System. *Introducing Teachbot. Virtual assistants for the classroom*. http://www.elzware.com/uk/teachbot.html.
10. FORLIZZI, J., ZIMMERMAN, J., MANCUSO, V., KWAK, S., 2007. *How Interface Agents Affect Interaction Between Humans and Computers*. New York, NY: ACM. Designing Pleasurable Products and Interfaces; 209–221.
11. HARRIS, R. A., 2005. *Voice interaction design : crafting the new conversational speech systems*. San Francisco, CA : Morgan Kaufmann.
12. HEISTERKAMP, P., 2001. *Linguatronic Product-Level Speech System for Mercedes-Benz Cars*. Morristown, NJ: Association for Computational Linguistics. Human Language Technology Conference; 1–2.
13. HOLM, N., NYLUND, S., 2009. *Rösten tar befälet*. Sydsvenskan, 2009-04-24, https://www.sydsvenskan.se/2009-04-23/rosten-tar-befalet.
14. IBM. Ideas from IBM. 5 innovations in the next 5 years. *Surfa med rösten: Att kunna prata med Internet och få svar*. http://www-05.ibm.com/se/ideasfromibm/five_in_five/?ca=se_five_in_five&me=w&met=se_hp_lead.
15. IKEA. *Ask Anna*. http://www.ikea.com/gb/en/.
16. JOKINEN, K., 2007. *Constructive Dialogue Management for Speech-based Interaction Systems*. New York, NY: ACM. International Conference on Intelligent User Interfaces; 22–22.
17. KUMAR, A., RAJPUT, N., CHAKRABORTY, D., AGARWAL, S. K., NANAVATI, A. A., 2007. *WWTW : The World Wide Telecom Web*. New York, NY: ACM. Applications, Technologies, Architectures, and Protocols for Computer Communication; Article No. 7.
18. KURNIAWAN, S. H., SPORKA, A. J., 2008. *Vocal Interaction*. New York, NY: ACM. Conference on Human Factors in Computing Systems; 2407–2410.
19. LARSEN, L. B., JENSEN, K. L., LARSEN, S., RASMUSSEN, M. H., 2007. *Affordance in mobile speech-based user interaction*. New York, NY: ACM. International Conference Proceeding Series; Vol. 309, 285-288.
20. LESTER, J., BRANTING, K., MOTT, B., 2004. *Conversational Agents*. In: Practical Handbook of Internet Computing, Singh, M. (Ed.). Baton Rouge, LA : Chapman Hall & CRC Press.
21. LOEBNER PRIZE IN ARTIFICIAL INTELLIGENCE. *The Loebner Prize in Artificial Intelligence. "The First Turing Test"*. http://www.aisb.org.uk/events/loebner-prize.
22. LOTSSON, A., 2009. *Sökmotorerna som förstår dig*. Computer Sweden, 2009-05-15, http://www.idg.se/2.1085/1.230050.
23. MCTEAR, M. F., 2002. *Spoken Dialogue Technology : Enabling the Conversational User Interface*. New York, NY: ACM. Computing Surveys; Vol. 34, No. 1, March 2002, 90–169.
24. NIELSEN, J., 2003. *Voice Interfaces: Assessing the Potential*. Jakob Nielsen’s Alertbox. http://www.useit.com/alertbox/20030127.html.
25. SAFFER, D., 2007. *Designing for interaction : creating smart applications and clever devices*. Berkeley, CA : New Riders.
26. SAFFER, D., 2009. *Designing gestural interfaces*. Sebastopol, CA : O'Reilly.
27. SHRIVER, S., TOTH, A., ZHU, X., RUDNICKY, A., ROSENFELD, R., 2001. *A Unified Design for Human-*

Machine Voice Interaction. New York, NY: ACM. Conference on Human Factors in Computing Systems; 247–248.

28. SIMONIN, J., CARBONELL, N., PELÉ, D., 2008. *Effectiveness and Usability of an Online Help Agent Embodied as a Talking Head*. New York, NY: ACM. International Conference on Multimodal Interfaces; 17–20.
29. SMARTHOME. Home Automation Superstore. *HAL 2000 Voice Control System*. http://www.smarthome.com/1450/HAL-2000-Voice-Control-System/p.aspx.
30. SMARTHOMEUSA.COM. The Smartest DIY Home Improvements. Home Automation, Security & Energy Savings. *Voice Control*. http://www.smarthomeusa.com/Shop/Voice.
31. SMITH, C., CHARLTON, D., ZHANG, L., CAVAZZA, M., HAKULINEN, J., TURUNEN, M., 2008. *An Embodied Conversational Agent as a Lifestyle Advisor*. Padgham, Parkes, Müller and Parsons (eds.) Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems. International Conference on Autonomous Agents; 1713–1714.
32. TÖRNBERG, U., 2009. *Smittspårning – direkt i din mobil*. Sydsvenskan, 2009-05-26, https://www.sydsvenskan.se/2009-05-25/smittsparning--direkt-i-din-mobil.
33. WAHLSTER, W., 2007. *SmartWeb : Multimodal Web Services on the Road*. New York, NY: ACM. International Multimedia Conference; 16–16.
34. WIKIPEDIA. *Ford Sync*. http://en.wikipedia.org/wiki/Ford_Sync.
35. WIKIPEDIA. *Natural Language*. http://en.wikipedia.org/wiki/Natural_language.
36. WIKIPEDIA. *Natural User Interface*. http://en.wikipedia.org/wiki/Natural_User_Interface.
37. WIKIPEDIA. *SixthSense*. http://en.wikipedia.org/wiki/SixthSense_(device).