Röststyrning – alla pratar om det

EXPERTARTIKEL

Av Cliff Ortmeyer, Premier Farnell och Farnell element14

Cliff Ortmeyer har arbetat i elektronikindustrin i 26 år med allt från produktutveckling till marknadsföring. Han har bland annat varit på ST Microelectronics och Coilcraft. Cliff började på Premier Farnell för sex år sedan och är idag globalt ansvarig för tekniks marknadsföring och utveckling av lösningar.

Så länge det har funnits datorer har ingenjörer letat efter bättre sätt för människor att interagera med tekniken. Fram till de senaste genombrotten med pekskärmar var de flesta gränssnitt mellan människor och datorer en avvägning mellan användarvänlighet och detaljstyrning. Pekskärmen blev ett mycket mer naturligt och intuitivt sätt att styra tekniken, något som till och med barn enkelt kan lära sig.

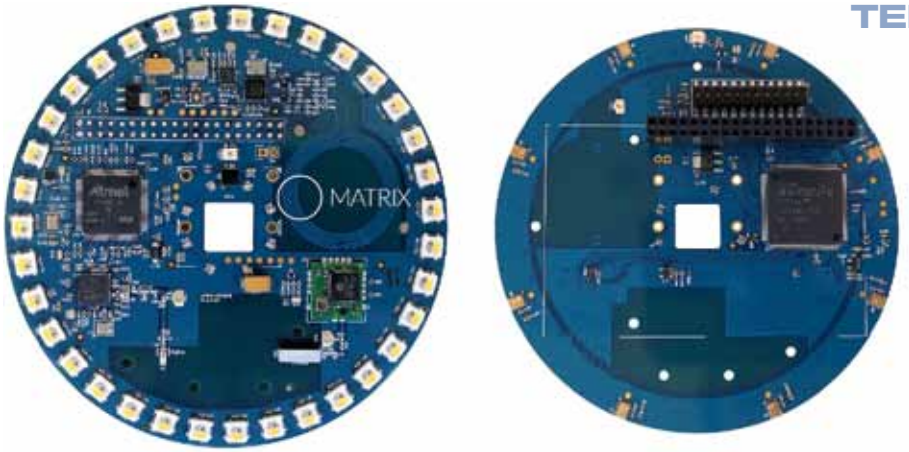
Pekskärmar passar dock inte i alla sammanhang. De gör enkla tillämpningar dyrare, de är opraktiska på små enheter, de är sårbara i installationer utomhus både när det gäller vandaler och väder, de kan utgöra säkerhetshot och dessutom måste användaren såklart befinna sig i närheten. I sådana här fall måste designern hitta ett sätt att intuitivt styra tekniken och samtidigt undvika de här nackdelarna. Då kan röstigenkänning och röststyrning vara den perfekta lösningen.

Att styra teknik via rösten är inte någon ny idé. Ända sedan de första datorerna byggdes har det gjorts försök att styra dem med tal. Framgångarna har varit varierade, men den senaste tiden har det gjorts stora framsteg inom algoritmer för taligenkänning. Dagens beräkningskraft har även gjort att utvecklarna kan ta fram teknik som snabbt och exakt kan svara på en mängd olika kommandon.

1952 gjordes det första försöket att skapa en dator som skulle förstå mänskligt tal, det skedde på Bell Laboratories. Systemet kallades Audrey. Det var väldigt enkelt och kunde bara förstå en del tal, och bara vissa personer.

TALIGENKÄNNINGEN FÖRBÄTTRADES i omgångar fram till 70-talet när amerikanska försvarsdepartementet drog igång Darpas SUR-program (Speech Understanding Research). Det var ett omfattande projekt som drevs mellan 1971 och 1976. Forskningen ledde så småningom till utvecklingen av Carnegie Mellons Harpy-system, som hade ett ordförråd på mer än tusen ord. Det här var inte bara en utveckling av tidigare system, i Harpy användes en helt ny typ av sökning som var mycket effektivare än tidigare. Med beam-sökning gick det att förutspå det ändliga nätverket av möjliga meningar.

På 80-talet gjordes stora framsteg inom taligenkänningen när HMM-modellen (Hidden Markov Model) utvecklades. Det här var en statistisk modelleringsteknik som kunde förutspå om enskilda ljud utgjorde ord. Det medförde att datorer kunde lära sig flera tusen ord. Nästa stora framsteg gjordes 1997,



Matrix Creator kan användas både som skal till Raspberry Pi och som fristående enhet. Kortet har sju Mems-mikrofoner som ger 360-graders ljudtäckning.

med det första systemet som förstod naturligt tal. Dragon Naturally Speaking kunde bearbeta ungefär 100 ord i minuten.

De här genombrotten banade väg för taligenkänningen. Det som behövdes för att göra tekniken mer tillgänglig var lägre kostnader och den beräkningskraft som krävs för att generera svar i realtid. Två jättar inom branschen gav oss nyligen det här – Google och Apple.

2011 lanserade Apple Siri, företagets smarta, digitala personliga assistent i iPhone 4S. Siri gav användaren en helt ny kontroll över systemet. Användarna kunde ringa sina vänner, diktera meddelanden och spela musik via röststyrning. Googles röstsökningsapp utvecklades 2012 och var ursprungligen tänkt för Apples iPhone. Där användes telefonens möjlighet att kommunicera med omvärlden till att jämföra sökfraser med data från användarsökningar som företaget samlat på sig i molnet. Möjligheten att jämföra med tidigare sökningar förbättrade noggrannheten avsevärt eftersom AI:n fick bättre förståelse för sökningens sammanhang. I praktiken var både Google Search och Siri sekundära gränssnitt vid sidan av pekskärmen. Amazon tog sedan konceptet till en helt ny nivå med Echo, där en digital, personlig assistent kombinerades med en högtalare helt utan pekskärm.

I takt med att de smarta, digitala, personliga assistenterna blev allt mer populära så har fler och fler designers och entusiaster börjat använda taligenkänning i sin utveckling. Det här säger även analytikerna på ABI Research,

som uppskattar att det kommer levereras 120 miljoner röststyrda enheter år 2021 och att röststyrningen kommer att vara det primära gränssnittet i folks smarta hem.

De designers som vill använda röststyrning i sina produkter har några saker de måste tänka på. Det går att skapa hela systemet från grunden och göra det möjligt att köra offline, men det skulle begränsa funktionaliteten avsevärt. Algoritmerna och biblioteken för taligenkänningen skulle begränsas av mängden minne, och det skulle vara svårt att införliva nya kommandon. Det går dock att göra. PocketSphinx har utvecklats för Android-enheter, och den senaste versionen kan användas som fristående app i enheter som kör Android Wear 2.0.

DAGENS UTVECKLARE vill dock kunna erbjuda en större mängd instruktioner, och då krävs en anslutning till molnet. De flesta molnleverantörer, som Amazon och Google, erbjuder taltjänster som är relativt billiga att bygga in i designen. Precis som med alla designbeslut så kommer din egen prioritetslista avgöra vilken tjänst som passar bäst. IBM erbjuder till exempel också en taltjänst på företagets plattform Watson Cloud. Plattformen är flexibel, och kan vara intressant om du vill dra nytta av IBM:s analytiska expertis snarare än att använda en mer allmän, kundfokuserad plattform.

Både Amazon och Google erbjuder plattformar som är skräddarsydda för marknaden med automation i hemmet. Bägge företagen har byggt upp ekosystem med några av de mest respekterade utvecklarna av produkter för hemmaautomation. Amazon samarbetar bland annat med Nexia, Philips Hue, Cree, Osram, Belkin och Samsung. Google har delvis liknande samarbeten som Amazon, bland annat med Hive, Nest, Nvidia, Philips Hue och Belkin.

De två företagen erbjuder tillgång till sina plattformar relativt billigt så att samarbetsföretagen ska kunna använda röststyrningstjänster från Amazon och Google i sina produkter. Med Amazons AVS (Alexa Voice Service) kan utvecklare integrera Alexa direkt i sina produkter. AVS innehåller även en komplett uppsättning resurser som API:er, SDK:er, utvecklingskit och dokumentation.

Google låter även utvecklare använda funktionerna i företagets digitala, personliga assistent Google Assistant via ett SDK. I Google Assistant SDK finns två sätt att integrera Assistant, Google Assistant-biblioteket och gRPC-API:t för Google Assistant. Google Assistant-biblioteket är skrivet i Python och kan användas på enheter med arkitekturerna linux-ARM v7l och linux-x86_64 (som Raspberry Pi 3 B och Ubuntu-datorer). Biblioteket är ett händelsebaserat högnivå-API som är enkelt att bygga ut. gRPC-API:t för Google Assistant är ett lågnivå-API. Du kan skapa kopplingar till det här gränssnittet i språk som Node.js, Go, C++ och Java på alla plattformar som har stöd för gRPC.

FÖR DEM SOM GÄRNA undviker de här tjänsterna och använder gränssnitt med öppen källkod finns andra alternativ. Mycroft är till exempel en kostnadsfri personlig assistent med öppen källkod för Linux-baserade operativsystem, där naturligt tal används i användargränssnittet. Mycroft är även modulärt så att användarna kan ändra komponenter i systemet. Jasper är ett annat alternativ med öppen källkod där utvecklare enkelt kan lägga till nya funktioner i programvaran.

Det passar bäst för kortdatorer som Raspberry Pi. Vissa nya kort har utvecklats särskilt för röststyrning, som Matrix Creator. Det kan användas både som skal till Raspberry Pi och som fristående enhet. Kortet har sju Mems-mikrofoner som ger 360-graders ljudtäckning. Det drivs av en Cortex M3 med 64 Mbit SDRAM-minne. Dessutom har det olika sensorer som gör att utvecklare kan lägga till funktioner. Kortet kan även användas med färdiga tjänster som Amazon AVS, Google Speech API och Houndify.

Mikrofonerna är såklart en viktig del av designen. Ofta används flera mikrofoner ordnade i en matris så att ljuden kan tas upp bättre och representeras mer verklighetstroget. Om tekniken för att sammanfoga ljuden från de olika mikrofonerna inte är inbyggd kan det behövas ytterligare designarbete och beräkningskraft. Dessutom är det oerhört viktigt med brusreducering så att instruktionerna tas emot felfritt.

Utvecklarnas mål är att skapa ett så intuitivt gränssnitt som möjligt mellan människa och maskin. Idag finns inga gränssnitt som går att jämföra med de instinktiva metoder som används i mänsklig kommunikation. Dagens röststyrning befinner sig på en nivå där processen känns nästan lika naturlig som att prata med en annan människa. Även om några av branschens största aktörer har tagit fram tekniken som används är den nu tillgänglig för utvecklare världen över. Eftersom mycket av bearbetningen görs i molnet krävs ofta mindre maskinvara än du tror. Dessutom finns specialanpassade kort, verktyg och tjänster som gör processerna mycket enklare, så idag kan du använda röststyrning inom alla typer av projekt. ■