VoiceXML : het Web leert spreken

Christophe Strobbe – Onderzoeksgroep Documentarchitecturen van de K.U.Leuven

INLEIDING

Op het einde van de jaren ’90 werd voorspeld dat de computer tegen het midden van dit decennium niet meer het belangrijkste toestel voor internettoegang zou zijn. Mobiele telefoons en PDA’s (personal digital assistants) kunnen nu reeds webtoepassingen gebruiken, maar ook huishoudtoestellen zoals de koelkast en de koffiezet zouden met het internet verbonden worden.
De broodrooster die telefonisch het weerbericht opvraagt en op basis daarvan een zonnetje of een regenwolk op je toast bakt is al een feit. De eerste webtoepassing voor de mobiele telefoon was het veel te vroeg gelanceerde WAP (Wireless Application Protocol).
De verwachtingen voor WAP waren hooggespannen: men verwachtte webpagina’s te zien zoals op een PC, maar men kreeg slecht monochrome ‘cards’ op een minuscuul schermpje. Bovendien verliep het ingeven van data zeer moeizaam, namelijk via de toetsen, terwijl de handigste interface van het toestel - de microfoon - niet eens gebruikt kon worden.
VoiceXML en ‘voice browsers’ kunnen hierin verandering brengen, maar ook deze technologie heeft zijn beperkingen.

Surfen op het web is voor de meeste gebruikers vooral een visuele ervaring. HTML-pagina’s bevatten tekst, afbeeldingen en links, maar slechts af en toe geluid. Input gebeurt met een muis en/of een toetsenbord, maar niet met spraak. VoiceXML kan hierin verandering brengen.
VoiceXML is een markuptaal die gesproken interactie tussen mens en computer mogelijk maakt. De taal werd oorspronkelijk ontwikkeld door het VoiceXML-Forum (VoiceXML 1.0), op basis van bestaande markuptalen uit de telefonie. Midden 2000 droeg het VoiceXML-Forum de verdere ontwikkeling van VoiceXML over aan het World Wide Web Consortium, dat ook HTML en XML ontwikkelde.
Hoewel VoiceXML gebaseerd is op technologie waar al ruim 20 jaar onderzoek naar verricht wordt, beginnen interactieve spraaktoepassingen nu pas door te dringen in de (Amerikaanse) bedrijfswereld.
VoiceXML is een markuptaal voor het beschrijven van dialogen die gebruik maken van spraakherkenning, spraaksynthese, digitale audio en DTMF-input (dat zijn de tonen die met de toetsen van de telefoon gemaakt worden). VoiceXML kan gezien worden als het samenkomen van twee verschillende ontwikkelingen.

Enerzijds zochten bedrijven in de telefoniesector naar een eenvoudiger manier om spraaktoepassingen te ontwikkelen; dit soort toepassingen zou dan een deel van het werk van een call center kunnen overnemen. Anderzijds werd gezocht naar technologie die het mogelijk moest maken om over het web te browsen met de telefoon. Er zijn verschillende redenen voor het samenkomen van deze trends. Telefoniebedrijven zoals Motorola hoopten dat het gebruik van webtechnologie het ontwikkelen van telefonietoepassingen gemakkelijker en goedkoper zou maken en ontwikkelden hiervoor markuptalen gebaseerd op XML. Ontwikkelaars van webtoepassingen en het World Wide Web Consortium (W3C) begonnen zich in deze markuptalen te interesseren omdat ze bruikbaar leken om webpagina's van spraak te voorzien.

INTERACTIEVE SPRAAKTOEPASSINGEN

De voice browser

Wat moet men zich nu voorstellen bij een spraaktoepassing op het web? Een gebruiker belt een bepaald telefoonnummer en wordt verbonden met een zogenaamde voice browser. Dit is een machine (hardware en software) die markuptalen voor gesproken interactie interpreteert om op basis daarvan gesproken output te genereren en gesproken input te interpreteren.

In systemen die op VoiceXML gebaseerd zijn, wordt de kern van de voice browser gevormd door de VoiceXML-interpreter. Andere belangrijke onderdelen zijn de componenten voor spraaksynthese (TTS: Text To Speech), voor de output van audiobestanden, voor het opnemen van spraak en voor spraakherkenning (ASR: Automatic Speech Recognition). De voice browser is verbonden met het telefoonnetwerk en met een klassieke webserver, vanwaar de nodige bestanden afgehaald worden.

Verschillen tussen webinterfaces en spraakinterfaces

VoiceXML (Voice Extensible Markup Language) is net als HTML (Hypertext Markup Language) een markuptaal en VoiceXML-bestanden worden net als HTML-bestanden op webservers opgeslagen. Beide talen dienen om webinformatie toegankelijk te maken maar doen dit op elk op een heel andere manier. Spraakinterfaces verschillen sterk van grafische interfaces en webinterfaces en er zijn dus ook een aantal belangrijke verschillen tussen HTML en VoiceXML:

– HTML beschrijft een tweedimensionale layout met visuele componenten zoals titels, paragrafen, afbeeldingen en formulieren, terwijl VoiceXML gebruik maakt van spraakherkenning en spraaksynthese en alleen de dimensie van de tijd kent.

– HTML dient voor het beschrijven van eenheden die uit een volledige pagina bestaan en die weergegeven worden door een webbrowser. De eenheden die VoiceXML beschrijft zijn dialogen die via de telefoon verlopen, en die zelf nog eens onderverdeeld kunnen worden in kleinere eenheden (forms en menu's).
– Een enkele HTML-pagina kan de gebruiker soms tientallen opties aanbieden, terwijl dit in spraaktoepassingen zeer gebruiksonvriendelijk zou zijn. In spraaktoepassingen probeert men het aantal opties in een bepaalde fase van een dialoog te beperken om het korte termijngeheugen van de gebruiker niet te overbelasten en om de prestaties van de spraakherkenning te vergroten.

Toepassingen

De dialogen die men met VoiceXML kan beschrijven, bestaan uit verschillende delen:

1. (eenvoudige) vragen met een beperkt aantal mogelijke antwoorden; de gebruiker kan antwoorden met zijn stem of met de toetsen van zijn telefoon;
2. antwoorden die de computer moet geven als de gebruiker niet antwoordt;
3. antwoorden die de computer moet geven als hij het antwoord van de gebruiker niet herkent;
4. dingen die de computer moet zeggen en/of doen als hij het antwoord van de gebruiker wel herkent.

Omdat dit soort dialogen grote beperkingen heeft, is VoiceXML niet geschikt voor alle soorten informatie die op het web te vinden zijn. VoiceXML wordt wel gebruikt voor toegang tot:

1. zakelijke informatie, b.v. telefonische bestellingen, home banking, aankomst- en vertrekuren van vliegtuigen;
2. publieke informatie, b.v. weer, verkeersinformatie, nieuws, beurskoersen;
3. persoonlijke informatie, b.v. kalenders, adres- en telefoonlijsten;
4. het zenden en ontvangen van e-mail per telefoon.

De hierboven beschreven toepassingen zijn webtoepassingen die men gebruikt via de telefoon, maar men wil VoiceXML ook inzetten op het "traditionele" Web, dat we benaderen via de computer. Hierop komen we later terug.

Soorten interactie

Er bestaan twee belangrijke soorten gebruikersinterfaces: door de gebruiker gestuurde (user-directed) en door de machine gestuurde (machine-directed). In het eerste type interface, stelt de gebruiker vragen of geeft hij commando's en de computer reageert; in het tweede type interface stelt de computer vragen en de gebruiker reageert. Sommige toepassingen wisselen af tussen beide soorten en worden

daarom mixed-initiative interfaces genoemd. In telefonietoepassingen spreekt men gewoonlijk over “directed dialogues” (dialogen die door de machine gestuurd worden) en “mixed initiative dialogues”. Directed dialogues zijn het eenvoudigst, omdat men de gebruiker om slechts één stukje informatie per keer vraagt. Het nadeel is dat ze omslachtig zijn, zoals het navigeren van een sterk vertakte menustructuur. De volgende dialoog geeft hiervan een voorbeeld: System: Please choose an application. The available choices are: banking, stocks, weather or travel. *User: Banking.* System: Welcome to Bank By Voice. You can access your checking, savings, or loan accounts, transfer funds or return to the main menu. *User: Savings.* System: You can hear about your savings balance, last five deposits, last five withdrawals or last five transactions. *User: Transactions*. ... Mixed initiative dialogues laten een natuurlijker gesprekspatroon toe. Het systeem stelt een complexe vraag, zodat de gebruiker verschillende stukjes informatie ineens kan geven. Dit soort dialogen is gebruiksvriendelijker, maar moeilijk om te ontwerpen en gevoeliger voor fouten. Het volgende stukje dialoog is een voorbeeld:
System: Welcome to Bank By Voice. What would you like to do?
User: Transfer a thousand dollars from checking to savings.
Een ander concept in spraaktoepassingen is “reusable dialogue components”: dit zijn vooraf geschreven scripts voor veel voorkomende stukjes dialoog, bijvoorbeeld het opvragen van het nummer van een kredietkaart. Dergelijke componenten kunnen ontworpen worden door experts in spraakinterfaces om ze daarna in verschillende toepassingen te hergebruiken. Hierdoor heeft men minder tijd nodig om een toepassing te ontwerpen en wordt de consistentie van de interface in de hele toepassing groter.

Ondersteunende markuptalen

VoiceXML beschrijft slechts het verloop van dialogen. Eenvoudig gesteld: als de gebruiker A antwoordt, wordt hij naar vraag X gestuurd, enzovoort. Om de input van een gebruiker te herkennen heeft men echter een andere markuptaal nodig, namelijk een markuptaal die definieert welke mogelijke antwoorden een gebruiker zou kunnen geven. Door de mogelijke antwoorden te beschrijven in een zogenaamde ‘grammar’, beperkt men het aantal woorden of patronen dat de spraakherkenningsmodule moet proberen herkennen. Dit is nodig omdat men de input van willekeurige gebruikers moet kunnen herkennen zonder dat de gebruiker de software moet trainen in het herkennen van zijn stem (zoals dit gebeurt bij dicteersoftware). De specificatie die het World Wide Web Consortium hiervoor ontwikkelt, heet

Speech Recognition Grammar Specification (SRGS).
VoiceXML beschrijft ook niet hoe een stukje dialoog door de spraaksynthesizer moet uitgesproken worden: snel of traag, met een hoge of een lage stem, enzovoort. Hiervoor ontwikkelt het W3C de Speech Synthesis Markup Language (SSML).
Telefonietoepassingen moeten ook gebruikers kunnen doorverbinden met een ander toestel en bijvoorbeeld telefoonconferenties mogelijk maken: hiervoor dient Call Control CCXML.

VOICEXML EN TOEGANKELIJKHEID

Toegankelijkheid in de VoiceXML-specificatie

VoiceXML komt uit een andere sector dan bijvoorbeeld de speech synthesizer DECTalk, die in de jaren '80 ontwikkeld werd door de Assistive Technology Group van Digital en heeft eigenlijk niets te maken met de hulpmiddelen voor spraaksynthese en spraakherkenning die door blinden gebruikt worden. De VoiceXML-specificatie hield oorspronkelijk dus weinig rekening met toegankelijkheid.
In een appendix van een vorige versie van de VoiceXML-specificatie verklaarde de Voice Browser Working Group hoe de richtlijnen voor toegankelijkheid van het Web Accessibility Initiative van toepassing zijn op VoiceXML:

1. VoiceXML dient alleen voor informatie gericht op auditieve interactie. Om dezelfde informatie toegankelijk te maken, moeten ontwikkelaars andere kanalen gebruiken die andere soorten interactie ondersteunen, bijvoorbeeld HTML of WML (voor WAP);
2. de Authoring Tool Accessibility Guidelines leggen uit hoe ontwikkelaars toegankelijke authoring tools kunnen ontwerpen. VoiceXML authoring tools worden in deze specificatie niet behandeld;
3. de User Agent Accessibility Guidelines leggen uit hoe ontwikkelaars toegankelijke user agents kunnen ontwerpen. Aangezien VoiceXML expliciet auditieve interactie en DTMF-input behandelt, zullen VoiceXML-user agents niet voldoen aan de User Agent Accessibility Guidelines.

Alfred Gilman van de Protocols and Formats Working Group (een subgroep van het WAI) leverde hierop commentaar. In de volgende versie van de specificatie, de Candidate Recommendation, heeft men de appendix over toegankelijkheid van VoiceXML aangepast:

1. terwijl VoiceXML in eerdere versies beperkt was tot spraaktoepassingen, worden spraaktoepassingen nu als de kern van het toepassingsgebied beschouwd. De meeste gebruikers zullen met VoiceXML-toepassingen interageren door te luisteren en te spreken, maar sommige gebruikers kunnen door

tijdelijke (of permanente) omstandigheden niet luisteren en/of spreken. Daarom moeten er andere manieren voorzien worden om te interageren met VoiceXML-toepassingen. Om de speciale software of hardware voor mensen met gehoor- of spraakproblemen te ondersteunen, voorziet de VoiceXML-specificatie de mogelijkheid om alternatieve tekst voor geluidsbestanden te gebruiken. Naast gesproken input zou het ook mogelijk moeten zijn om input van een computerklavier te gebruiken;

2. de VoiceXML-specificatie erkent nu dat de Web Content Accessibility Guidelines 1.0, de Authoring Tool Accessibility Guidelines 1.0, de User Agent Accessibility Guidelines 1.0 en de XML Accessibility Guidelines ook van toepassing zijn voor VoiceXML. Wat de toegankelijkheid betreft van software om VoiceXML-toepassingen te ontwikkelen, is het eerder twijfelachtig dat deze binnen afzienbare tijd toegankelijk wordt. Midden 2002 bleken de meeste tools en alle webgebaseerde ontwikkelomgevingen voor VoiceXML zeer slecht toegankelijk of helemaal ontoegankelijk te zijn. Omdat de markt voor VoiceXML-software nog vrijwel onbestaand is, zijn de aanbieders van deze software waarschijnlijk niet gemotiveerd om hier snel iets aan te veranderen;
3. tenslotte geeft de appendix nog enkele bijkomende richtlijnen voor de toegankelijkheid van VoiceXML toepassingen.
   - hergebruik gebruiksvriendelijke navigatiestructuren in verschillende applicaties, bijvoorbeeld de navigatietechnieken uit de Amerikaanse standaard voor Digital Talking Books (ANSI/NISO Z39.86-2002, zie http://www.loc.gov/nls/z3986/) en ook het artikel Daisy for Dummies in dit nummer van IM;
   - elk element waarin een ‘event’ kan optreden zou ook ‘catch’-elementen moeten definiëren, waaronder één die de mogelijkheid biedt om een fout recht te zetten;
   - stel gebruikers in staat om de tijdspanne voor een ‘timeout’, de snelheid van synthetische spraak en andere variabelen te beïnvloeden die ervoor zorgen dat men voldoende tijd krijgt om te reageren of om op een vraag te antwoorden. Dit is vooral belangrijk wanneer gedetecteerd wordt dat de gebruiker een hulpmiddel gebruikt in plaats van te spreken of te luisteren. De extra tijd is vooral nuttig voor gebruikers met cognitieve beperkingen;
   - vestig de aandacht op alternatieve methodes om van een equivalente dienst gebruik te maken, waaronder doorverbinden met een operator, teksttelefoon, enzovoort, of de beschikbaarheid van dezelfde informatie op het web.

Ondertussen schijnt de Voice Browser Working Group van het World Wide Web Consortium reeds te werken aan een document over accessibility requirements voor VoiceXML 3.0, maar dit document is nog niet publiek.

Het EU-project VISUAL

VoiceXML wordt tot nu toe alleen gebruikt in telefonietoepassingen en niet op het traditionele web.
Men zoekt wel naar manieren om VoiceXML of een gelijkaardige markuptaal te combineren met HTML om spraaksynthese en spraakherkenning te integreren in de web user interface. Het Europese onderzoeksproject VISUAL, waarin ook de onderzoeksgroep DocArch van de K.U.Leuven betrokken is, onderzoekt onder andere hoe VoiceXML gebruikt kan worden om webpagina's toegankelijker te maken voor blinden en slechtzienden, en vooral in de context van e-learning. Hierbij wordt gebruik gemaakt van ConPalabras, een browser-plugin die eenvoudige VoiceXML-documenten kan lezen. Deze plug-in werkt enkel met Internet Explorer 4 of recenter en is gratis beschikbaar op www.conpalabras.com. Naast een recente versie van Internet Explorer heeft men ook een minstens even recente Java Runtime Environment nodig. De plug-in zelf doet spraaksynthese, maar wie ook spraakherkenning wil gebruiken heeft ook nog ViaVoice van IBM nodig. Voorlopig zijn er nog niet zoveel websites waar men ConPalabras kan gebruiken: er zijn enkele demonstratiepagina's op de ConPalabras-website en er komen een aantal voorbeelden op de e-learningsite van VISUAL. De manier waarop HTML en VoiceXML gecombineerd worden in webpagina's die 'geprepareerd' zijn voor ConPalabras is echter niet elegant: de HTML-pagina moet een stukje JavaScript bevatten met een verwijzing naar een VoiceXML-pagina; deze URL wordt via JavaScript doorgegeven aan een Java-applet dat zorgt voor de interactie met de eigenlijke plug-in. De plug-in zelf ondersteunt slechts een kleine subset van VoiceXML; de dialogen die men kan creëren zijn dus niet vergelijkbaar met wat men hoort in telefonietoepassingen.
Het combineren van spraakinput en -output met traditionele webtechnieken is echter heel wat complexer dan het creëren van 'zuivere' telefonie- of webtoepassingen. Dit soort toepassingen hebben 'multimodale interfaces' omdat ze verschillende input- en outputmodaliteiten combineren: enerzijds spraak en geluid (VoiceXML) en anderzijds tekst (HTML).

Multimodale interactie

Het doel van het onderzoek naar multimodale interfaces is het integreren van verschillende communicatiemethodes: gesproken interactie, gebaren, aanraking, enzovoort. Met de specificaties en de

technologie die nu bestaan, staat de combinatie van gesproken interactie met de klassieke PC-interactie (toetsenbord voor input, tekstuele en visuele output) nog het dichtst bij een doorbraak. De Voice Browser Working Group van het World Wide Web Consortium formuleerde in 2000 reeds enkele vereisten in verband met de multimodale aspecten van markuptalen voor gesproken interactie, maar dit werk werd begin 2002 overgedragen op de nieuwe Multimodal Interaction Activity. Deze werkgroep ontving reeds verschillende voorstellen voor markuptalen van bedrijven en industriële consortia. SALT (Speech Application Language Tags) is een voorstel van het SALT-Forum, dat van mening is dat VoiceXML niet geschikt is voor multimodale toepassingen. Microsoft, de belangrijkste speler in het SALT-Forum, werkt aan software waarmee spraak kan geïntegreerd worden in webtoepassingen. Niet iedereen gelooft dat VoiceXML ongeschikt is voor multimodale toepassingen: in november 2001 dienden IBM, Motorola en Opera een voorstel in dat toont hoe XHTML en VoiceXML gecombineerd kunnen worden: XHTML+Voice Profile. IBM stelde eind januari 2003 reeds een beta-versie van een Multimodal Browser and Toolkit ter beschikking, waarmee men multimodale toepassingen kan gebruiken die gebaseerd zijn op XHTML +Voice. Meer informatie vindt men op http://www.alphaworks.ibm.com/tech/mmb.

Besluit

Hoewel spraaksynthese en spraakherkenning geen nieuwe technologieën zijn, zijn VoiceXML en aanverwante standaarden nog niet relevant voor blinde of slechtziende computergebruikers. De introductie van multimodale toepassingen kan hierin wel verandering brengen. Microsoft en IBM, die elk hun eigen technologie tot een standaard voor multimodale applicaties proberen te verheffen, hebben gelukkig heel wat ervaring op het gebied van toegankelijkheid. Hopelijk zullen ze bij het ontwikkelen van deze nieuwe technologie rekening houden met ‘design for all’.

Nota: De elektronische versie van dit artikel bevat een groot aantal aanklikbare hyperlinks.