Stemherkenning en spraaksynthese: geen gemakkelijke wetenschap

In ons decembernummer van vorig jaar kondigden we het Anty-robotproject voor gehospitaliseerde kinderen aan. Met hulp van het Brussels Hoofdstedelijk Gewest ontwikkelt de onderzoeksgroep Robotics and Multibody Mechanics (R&MM) van de VUB tegen 2007 een eerste prototype. Maar dit project heeft méér facetten dan de "robot" alleen. Een knuffelrobot moet tot op zeker niveau "sensoring" hebben (zien, voelen). En natuurlijk heeft Anty een stem nodig om te communiceren met en te reageren op de patiëntjes.

Voor deze "stem"-aspecten wordt de Anty-onderzoeksgroep spontaan bijgesprongen door de onderzoeksgroep DSSP (Laboratorium voor Digitale Spraak en Audio Processing), een afdeling van het departement ETRO (Elektronica & Informatica). Ten gepaste tijde zal Anty "beroep" kunnen doen op resultaten van fundamenteel en toegepast onderzoek van deze groep. Daarenboven zijn er nu al eindwerken gestart om Anty een eigen stem en een (brabbel)taaltje te geven (stemsynthese). De eindwerken hebben ook als doel om deze stem uit te rusten met een specifieke intonatie gericht op de stemming van het gehospitaliseerde kind (afgeleid uit hoe het kind iets zegt, dus uit stemherkenning). *Industrie Technisch Management* sprak over spraakonderzoek met **prof. dr. ir. Werner Verhelst**, hoofd van het Laboratorium voor Digitale Spraak en Audio Processing aan de **VUB**.

DIGITALE SPRAAK EN AUDIO

Digitale spraak heeft twee kanten: verstaan (spraakherkenning) en spreken (spraaksynthese). In België heeft men bij dit onderwerp meteen de neiging om te verwijzen naar het gekende Ieperse debacle, maar alle universiteiten werken weldegelijk zeer fundamenteel aan beide aspecten. En er zijn ook, reeds toepassingen waar deze technologie sociaal meer is dan een gadget.

Zo is er een SBO (strategisch basis onderzoek)-project, een door het IWT gefinancierde samenwerking tussen VUB, RUG, KU-Leuven en UA waarbij gewerkt wordt aan een combinatie van spraakherkenning en spraaksynthese om mensen met dysfasie (uitspraakmoeilijkheden) en dyslexie (met bijhorende leesproblemen) te ondersteunen in hun oefeningen. Men stelt namelijk vast dat er enerzijds te weinig logopedisten zijn om probleemkinderen op te vangen. Anderzijds hebben ouders te weinig tijd (en ook niet steeds het geduld) om hun kinderen veelvuldig te laten oefenen. En er is de "gewenning" bij de hulpverleners die dan denken dat het verbetert en de oefeningen verzwaren, met als resultaat dat het kind ontmoedigd vastloopt.

Een computer kan door het kind of de volwassenen met de leeshandicap zelf (en "neutraal", "zonder dwang") worden gebruikt en heeft het bijkomende voordeel dat het steeds een objectieve waarnemer is. Dit vergt een computer die kan herkennen wat de persoon in kwestie leest en die dit ook kan corrigeren. En niet voor één oefening, maar met een gedifferentieerd aanbod oefeningen. Vandaag wordt voor het verbeterend voorlezen van de oefeningen gewerkt met bestaande vaste opnames. Ideaal is echter te werken vanuit geschreven teksten die door de computer herkend en gedebiteerd kunnen worden.


RV/VUB

Natuurlijk heeft Anty (hier op het autosalon) een stem nodig om te communiceren met de patiëntjes.

In het luik "herkenning" is men vandaag ver in "geoefende woorden herkennen". Dat betekent dat de computer gesproken woorden van één of een beperkt aantal personen, na een "leerproces" (waarin deze personen de woorden een groot aantal keren uitspreken), kan herkennen. Er wordt gewerkt aan "zinsherkenning" (bijv. het uitschrijven van gedicteerde brieven, van interviews en van spontane spraak in het algemeen). Omdat deze systemen zijn gebaseerd op voorgeprogrammeerde grammatica, wordt het voor de computer echter al moeilijk als de persoon hapert in zijn zin, zich herpakt binnen de zin. En wie doet dat niet? In dit project gaat het nog een stap verder: de computer moet over "artificiële intelligentie" beschikken om woorden die gehakkeld of verkeerd worden gelezen, toch te herkennen om ze dan te corrigeren en juist voor te zeggen.

In het luik van de VUB, spraaksynthese, wil men "natuurlijke" (ook Vlaams klinkende) spraak realiseren vanuit "geschreven teksten". Het lijkt evident, maar vandaag lukt men er in om korte teksten te genereren met een vrij egale (computer)stem of korte meer natuurlijk klinkende boodschappen binnen een smal toepassingsgebied. De voorbeelden

kent u: "het door u gevraagde nummer is...", "u moet nu rechts afslaan"... En zelfs in deze eenvoudige toepassingen is het evenmin foutloos. Dit onderzoek wil een stap voorwaarts zetten met enerzijds het luik "gesproken instructies/helpsystemen" (want het gaat om mensen met leesproblemen, dus een gedrukte handleiding zal niet veel helpen), anderzijds "het voorlezen van de oefentekst" en "het verbeteren van fout gelezen oefenteksten". Dat vergt een zeer hoge spraakkwaliteit, want het gesprokene heeft een voorbeeldfunctie. De enige "vereenvoudiging" is dat men kan werken met een (ver)eenvoudig(d)e woordenschat in een taal zonder complexe Latijnse zinsconstructies. Ook binnen het Anty-project zal "hoge kwaliteit spraaksynthese" nodig zijn, wil men de knuffelrobot een "informatieverstrekkende rol" geven of hem "spontaan zinnige teksten laten antwoorden".

EVOLUTIE IN SPRAAKONDERZOEK

Er moet dus nog heel wat strategisch en fundamenteel onderzoek gebeuren vooraleer men zo ver zal zijn. Nochtans heeft het spraaksyntheseonderzoek al een hele weg afgelegd. En al zou men het niet denken, de onderliggende problemen van spraaksynthese zijn gelijkaardig en even complex als die in de spraakherkenning.

In spraakherkenning is er een probleem dat in spraaksynthese niet bestaat: het "achtergrondgeluid" dat het moeilijk maakt om wat moet herkend worden voldoende zuiver te detecteren. Problemen betreffende intonatie, klemtonen, het fenomeen van aan elkaar gekoppelde klanken in een vlot gesproken zin, zijn voor beiden moeilijk. Voor spraakherkenning werkt men met fonemen (klanken) die worden samengebracht tot voorspelbare woorden. Voor spraaksynthese moet men de juiste fonemen kiezen om het woord in die specifieke zin correct uit te spreken. Beiden werken op basis van een "taalmodel" om te kijken welke woorden achter elkaar kunnen staan, want de woordopeenvolging in een "dialoog" verschilt van de woordvolgorde in een "voordracht" en is nog eens anders bij het "bestellen van een treinticket". Het moeilijkste in spraakherkenning blijkt het uitschrijven van een "interview" te zijn. (Vinden journalisten trouwens ook), want in herkenning van spreektaal kampt men met het probleem dat de mens zichzelf soms herhaalt, pauzes neemt, de klemtoon legt op het hoofdwoord, enz. In computerstemsynthese wil men dergelijk vlot spreekgedrag nabootsen, en ook dat is zeker niet eenvoudig.

Selma Yilmazyildiz met een maquette van Anty omringd door medewerkers van het Anty project (vlnr: Werner Verhelst, Ivan Hermans, Kristof Goris, Jelle Saldien en Lukas Latacz) in de anechoische kamer.

SPRAAK IS MEER DAN TEKSTINHOUD

In de loop van de studies over digitale spraak is men voor de praktijk afgestapt van de zeer fundamentele optie van de articulatorische synthese. Een "fysieke simulatie" van de mond, stembanden, neus om te komen tot een "spraakmachine" is helemaal niet gemakkelijk, vandaag zelfs niet realistisch: hoe de werking van alle spieren in het hoofd opmeten om van daaruit "spreken" te simuleren? Men heeft de traagheden van de spierreacties, die de klankholte beïnvloeden waardoor zinnen in één geheel worden geproduceerd, waardoor letters heel anders worden uitgesproken in de ene samenstelling dan de andere. Neem een "a" in "bal" of in "kat". Die kan dus dof of helder klinken. En terugrekenen vanuit "de klankenstroom" is zeer moeilijk omdat gelijkaardige klanken op uiteenlopende manieren kunnen geproduceerd worden.

Daarom gaat men vandaag in toepassingen uit van een meer pragmatische benadering op basis van databases van gekende fonemen, woorden en zelfs stukken van zinnen. Voor spraakherkenning wordt getracht om deze stukken te herkennen. Bij spraaksynthese worden ze via een "filtering" aan elkaar gekoppeld om de computerstem "zo natuurlijk mogelijk" te laten klinken. Doordat de database van mogelijkheden steeds groeit en de computers steeds krachtiger worden, komt men stilaan naar "aanvaardbare" herkenning en niet te storende computerstemmen, tenminste voor "smalle toepassingen".
Naast het juist samenstellen (of herkennen) van de tekst, kampen onderzoekers in "digitale spraak", dus zowel in herkenning als synthese, met het probleem dat "de tekst" maar een onderdeel is van "de gesproken taal". Tekst is de woordenschat en de inhoud. Dat is in spraaksynthese al moeilijk genoeg, want hoe zegt de computer de geschreven zin: "hij kwam brood bedelen"? Brengt hij brood of vraagt hij brood? En 8080 in "een *Intel 8080*" (Intel tachtig tachtig) of in "de kostprijs is 8080 euro" (achtduizend tachtig). Naast woordenschat is er tevens "intonatie" en "emotie". Iemand kan woordelijk iets zeggen, maar via de intonatie van de stem en de context van het gesprek, weet de toehoorder dat het tegenovergestelde bedoeld werd: "en gij gelooft dat!?".

In spraaksynthese was men al lang blij als de computer zijn boodschap "op vlakke wijze" kon zeggen. Ook in spraakherkenning is dit luik lang niet zo sterk onderzocht als de eigenlijke "herkenning van de inhoud". Nu zijn er wél economisch belangrijke toepassingen waarvoor onderzoek betreffende "emotieherkenning" belangrijk is. Een eerste voorbeeld is het controleren op basis van het gedrag van de stem tijdens een gsm-gesprek of een truckchauffeur vermoeid is of niet. Een ander voorbeeld, en een vraag van de industrie, is de controle op computergebaseerde telefoongesprekken. De computer verstaat de klant niet altijd vanaf de eerste maal en die kan geërgerd geraken waardoor hij voor de computer helemaal onbegrijpbaar wordt. Maar de computer blijft natuurlijk "bevestiging vragen". Neem aan dat je telefonisch een vliegticket wil bestellen en je wil naar "Amsterdam" en de computer gokt op "Atlanta", een tweede maal opnieuw en dan zeg je "verdomme" en hij vraagt "Rome" en ga zo maar verder. Niemand wil zijn klant verliezen door een stomme computer. Dus dergelijk "emotioneel stemgedrag" moet - los van de woordelijke boodschap - kunnen worden gedetecteerd, want dan moet een menselijke operator overnemen en de brokken lijmen. Hetzelfde

kan gelden in callcenters. Aan de VUB wordt er ook op het vlak van "emotieherkenning" onderzoek verricht en van dat onderzoek zal ook het Anty-project kunnen profiteren.

SYNCHRONISATIE EN SAMENSPRAAK

Een vandaag - voor de onderzoeksgroep van de VUB althans - realiseerbare toepassing is het synchroniseren van twee stemmen die hetzelfde zeggen, het desgewenst vervangen van de spraak van de ene met behoud van een aantal aspecten van de stem van de andere, enz.

Een praktisch voorbeeld is de karaoketoepassing. Hiermee kan men bij een amateur die vals zingt zijn stemfrequentie, ritme, de ademhalingspauzes... vervangen door die van de oorspronkelijke zanger met echter het behoud van het timbre (klankkleur van de stem) van de amateur. Dan lijkt het of de valszinger toch goed kan zingen. Men kan op gelijkaardige wijze speelfilms synchroniseren (spraak van buitenopnames vervangen door in de studio ingesproken teksten, zonder het huidige probleem dat "de mondbeweging" en "de spraak" niet overeenkomen). Het laat ook toe om vertaalde spraak beter te synchroniseren met de mondbeweging. Of de stem van de dubber perfect kunnen doen lijken op die van de oorspronkelijke acteur (al lijken de meestal lokale acteurs dat eigenlijk niet te willen). Men kan vanuit één opname een meerstemmig koor realiseren. En ook hier weer: alhoewel het onderzoek niet gebeurt in functie van het Anty-project, kan men zich inbeelden dat deze techniek toch wordt geïntegreerd: misschien kan Anty zo meezingen met het gehospitaliseerde kind?

HET ANTY-EINDWERK

Naast dit "onderzoek waarvan de resultaten later kunnen worden overgezet", is men binnen DSSP ook gestart met specifiek op Anty gericht onderzoek. In tegenstelling tot het Anty-robotproject van *Robotics and Multibody Mechanics* (R&MM) is men binnen DSSP begonnen zonder specifiek hiervoor beschikbare fondsen. Het onderzoek gebeurt via eindwerken.

Een eerste eindwerk loopt dit jaar al en gaat over "een stem voor Anty". Omdat "onbemand met zinnige teksten antwoorden" vandaag nog te ver in de researchfase zit, en een stomme robot geen echt communicatief gegeven is, heeft men een andere benadering bedacht. Anty zal voor een aantal functies werken als een "afstandsbediende pop", maar daarnaast wil men het kind "spontane reacties" van de knuffelpop geven. Daarvoor krijgt Anty een "brabbeltaaltje" waarmee de knuffel "spontaan" kan inspelen op de gevoelens die het kind uit. En dat is wel een bereikbaar project. Vanuit een database met opnames van de professionele radiostem van Ivan Hermans, de geestelijke vader van Anty (niet dat dit technisch gesproken zijn stem moest zijn, maar het is zeker symbolisch leuk), worden via stemsynthese klanken en woorddelen gehaald en samengesteld tot een soort kindertaal. Het specifieke hierbij is dat bepaalde intonaties die een "gevoel" opwekken via algoritmes worden gerealiseerd.
Mensen herkennen emoties, zelfs als men de taal niet verstaat. Het gaat om toonhoogtes, snelheid van uitspreken, stemkwaliteit (bijv. met een krop in de keel)... die samen een aantal basisintonaties vormen. Men heeft "verbiedend", "berispend" en "aandacht trekkend" (een vlakke computerstem die zegt "het brandt en u wordt verzocht het gebouw te verlaten" zal waarschijnlijk minder effect uitlokken dan iemand die gilt "brand, brand!"). Dan is er "informerend" (goed voor in een later stadium in combinatie met beeld, als Anty wordt ingezet om dingen zoals "je wordt geopereerd" uit te leggen). Direct nuttige emotionele stemmen zijn: "troostend" en "bemoedigend", ook wel "goedkeurend" of "verbaasd" (bijv. in combinatie met "visie": als er een "bekend" iemand binnenkomt).

De "relevante" kenmerken detecteren, extraheren en in de "Anty-taal" integreren zou binnen het kader van één, desgewenst met een vervolgeindwerk moeten kunnen worden gerealiseerd. Zo zou Anty een stem krijgen. De computerpower moet niet in Anty zitten (die is batterijgestuurd, moet dus zo licht mogelijk zijn). De robot moet wel beschikken over een betrouwbare communicatie naar de server. Het resultaat van de synthese doorsturen vergt geen computervermogen. Het in real-time genereren van emotionele taal wel, maar dat is een probleem van de computerpower op de server. ■

VUB ETRO

Bij de VUB is Elektronica & Informatica (terug) één departement. Dit heeft te maken met de synergie tussen beiden: informatica steunt op elektronica en daarenboven is elektronica-ontwikkeling grotendeels gebaseerd op programmatie. Vooral omdat de VUB zich specialiseert in fundamenteel onderzoek van signaalverwerking. En eigenlijk is het samengaan van beide specialiteiten nog fundamenteler want zowel informatica als elektronica (zelfs de hardware-ontwikkeling) steunt op "wiskunde", het ontwikkelen van gepaste algoritmes die eens geïmplementeerd of in software of in hardware-opbouw doen wat men binnen de applicatie wil dat ze doen.

Het departement Elektronica & Informatica (ETRO) is opgedeeld in een aantal onderzoeksgroepen. Het onderzoek naar "elektronica-hardware" zit meer in LAMI (chips en sensoren). De grootste groep is IRIS (beeldverwerking). Dan is er nog TELE (telecom, de echte informatica dus). En de jongste telg is Beeld & Geluid (AVSP: Audiovisuele signaalprocessing), een multidisciplinaire groep met mensen van IRIS en DSSP. Natuurlijk is er ook DSSP, digitale spraak en audio processing. Deze discipline is eigenlijk zeer recent. Digitale audio en spraaktechnologie zijn (in België) gestart in de zeventiger jaren (aan de universiteiten van Gent en Bergen). Aan de VUB is het onderzoek gestart in de jaren '80 met het doctoraat van prof. Verhelst die een groep uitbouwde van een tiental onderzoekers en studenten. Het gaat eerder om "strategisch onderzoek": vanuit praktische vraagstellingen rond datapatronen van "spraak" komen tot bruikbare toepassingen en fundamentele inzichten.

Lukas Latacz van het Anty-project en eindwerkstudente Selma Yilmazyildiz (studie master of applied computer science, ir dept. VUB) doen een soundcheck voor opname.