

Bel dan James, onze huiscomputer!

Toekomst of realiteit?

Inleiding

Het is plotseling gaan sneeuwen en Monique staat in een lange file. Ze kijkt zenuwachtig op de klok. Opeens roept zij “*bel naar huis*”, en de autotelefoon kiest automatisch het juiste nummer. De telefoon gaat tien keer over, maar niemand neemt op. “*Vreemd*”, denkt ze. “*Bel dan James, onze huiscomputer!*” “*Met James.*” “*Zet de verwarming om acht uur aan, en neem vanavond de film Casablanca op.*” Als ze om negen uur thuiskomt doet haar man open en zegt tevreden: “*Hoi, ik ben ook net thuis. Wat goed dat je de verwarming en de videorecorder al aangezet hebt.*”

In het bovenstaande verhaal worden twee toepassingen van taal- en spraaktechnologie gebruikt, namelijk een ‘handen-vrij telefoon’ en een ‘gesproken-dialoogsysteem’. “*Is dit toekomst of realiteit?*”, zult u zich wellicht afvragen. Wel, na lezing van dit artikel weet u het antwoord op deze vraag. In dit artikel zal ik eerst een beschrijving geven van een gesproken-dialoogsysteem. Vervolgens wordt uitgelegd hoe automatische spraakherkenning werkt, en enkele toepassingen worden genoemd.

Openbaar Vervoer Informatie Systeem

Op de afdeling Taal en Spraak van de Katholieke Universiteit Nijmegen wordt gewerkt aan de ontwikkeling van een gesproken-dialoogsysteem voor het Nederlands. Als concrete toepassing is in eerste instantie gekozen voor een Openbaar Vervoer Informatie Systeem (OVIS) dat automatisch via de telefoon het grootste gedeelte van de informatie kan verstrekken die in het spoorboekje van de Nederlandse spoorwegen (NS) te vinden is. Veel mensen zijn bekend met het reisinformatieprogramma van de NS (vaak reisplanner genoemd) dat beschikbaar is op floppy en op WWW. Bij dit programma moet je de vragen intypen en krijg je de antwoorden op je beeldscherm te zien. OVIS is de ‘gesproken versie’ van deze reisplanner omdat je tegen het systeem kunt spreken en de antwoorden te horen krijgt. Omdat je geen toetsenbord en beeldscherm meer nodig hebt, kun je OVIS simpelweg bellen met iedere telefoon.

De architectuur van OVIS is te zien in figuur 1. De database is vrijwel dezelfde als die van de zojuist genoemde reisplanner. Behalve deze database en een telefooninterface bestaat OVIS uit vier componenten. Hier volgt een voorbeeld van een fictieve dialoog met OVIS. Stel dat iemand OVIS opbelt.

OVIS: “*Van welk station naar welk station wilt u reizen?*”

De openingszin is belangrijk, want als gewoon ‘goedemorgen’ gezegd wordt, zijn mensen geneigd om allerlei lange verhalen te gaan vertellen. Als er echter begonnen wordt met een gerichte vraag, geven de bellers vaker een bondig antwoord zoals bijvoorbeeld:

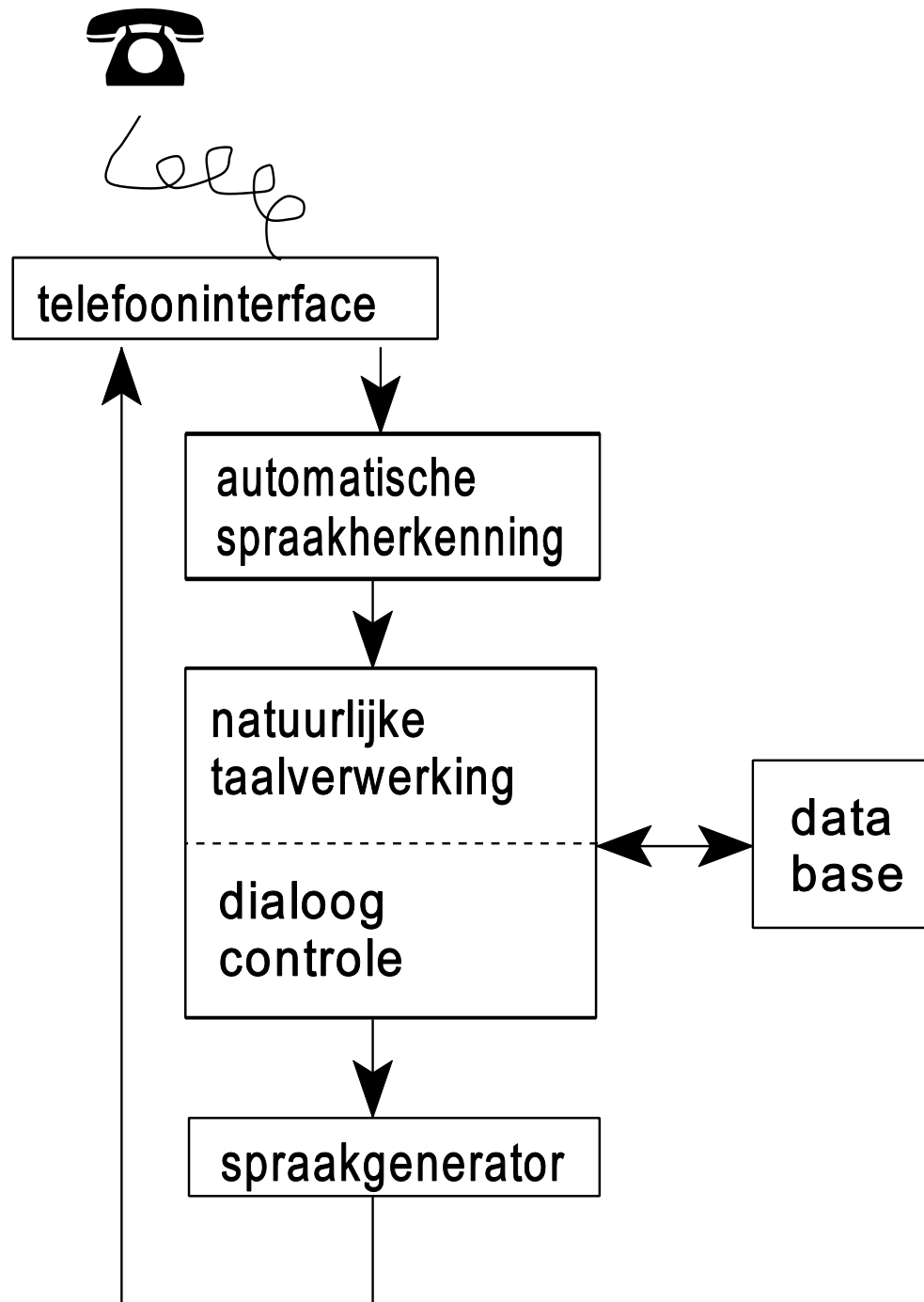
beller: “*Ik wil morgen van Nijmegen naar Tilburg.*”

De spraakherkenner zet het binnenkomende spraaksignaal om in een reeks woorden. De natuurlijke taalverwerkingsmodule zoekt in de herkende woorden naar concepten, d.w.z. voor de applicatie belangrijke informatie eenheden zoals stationsnamen en tijdstip van vertrek of aankomst. In dit geval zal hij de concepten ‘morgen’, ‘Nijmegen’ en ‘Tilburg’ vinden. De dialoog-controlemodule slaat deze informatie op en kijkt of er nog iets ontbreekt. Bijvoorbeeld in het bovengenoemde antwoord ontbreekt tijdstip van vertrek of aankomst. Omdat de door de beller gegeven informatie nog niet compleet is, zal het systeem om aanvullende gegevens moeten vragen. De dialoog-controlemodule formuleert de vraag (in tekstvorm). De spraakgenerator zet deze vraag (tekst) om in spraak. Deze gesproken vraag wordt via de telefoon naar de gebruiker gestuurd.

OVIS: “*Hoe laat wilt u morgen van Nijmegen naar Tilburg reizen?*”

beller: “*Ik wil morgenavond om 8 uur vertrekken.*”

De spraakherkenner herkent weer woorden, en de natuurlijke taalverwerkingsmodule zoekt naar concepten. In dit geval ‘morgenavond’ en ‘om 8 uur’. Ofschoon eerder al ‘morgen’ herkend was, is het concept ‘morgenavond’ toch nog belangrijk, namelijk om te weten of de gebruiker om 8 uur ’s morgens of ’s avonds wil vertrekken. De dialoog-controlemodule ziet dat de vraag nu compleet gespecificeerd is, zoekt het antwoord op in de database en formuleert het antwoord (in tekstvorm). Dit antwoord wordt met de spraakgenerator weer in spraak omgezet en naar de gebruiker gestuurd. Tenslotte vraagt het systeem of de gebruiker nog andere informatie wil. Als de gebruiker ontkennend antwoordt, bedankt het systeem voor het gebruik en wordt de verbinding verbroken.



Figuur 1. Architectuur van het Openbaar Vervoer Informatie Systeem.

Na uitgebreide tests is OVIS inmiddels in gebruik genomen. Wie Openbaar Vervoer Reisinformatie belt (0900-9292), krijgt meestal een persoon van vlees en bloed aan de lijn. Echter van middernacht tot 6 uur 's ochtends en bij wachttijden van meer dan 30 seconden krijgt men OVIS aan de lijn. In het derde kwartaal van 1997 werden van de 2.4 miljoen gesprekken er 200.000 automatisch verwerkt door OVIS. Hierdoor daalde het aantal mensen dat de in-gesprektoon te horen kreeg van 26% naar 17%. Een belangrijk onderdeel van OVIS is spraakherkenning. Hoe dat in zijn werk gaat, wordt hieronder uitgelegd.

Automatische spraakherkenning

De eerste serie spraakherkenners sloegen voor ieder woord dat herkend moest worden een of meerdere voorbeelden op. Voor een binnenkomend (onbekend) woord bepaalden ze de gelijkenis met de opgeslagen voorbeelden, en selecteerden het voorbeeld met de grootste gelijkenis. Deze techniek werkt alleen goed als het aantal te herkennen woorden klein is. Verder moesten de woorden los uitgesproken worden, met een pauze na ieder woord, omdat deze spraakherkenners er anders vaak niet in slaagden om begin en einde van de woorden te vinden. Voor vele hedendaagse applicaties is het echter noodzakelijk dat tienduizenden woorden herkend worden die natuurlijk (dat wil zeggen zonder pauzes) uitgesproken zijn. Bijvoorbeeld OVIS kan ruim 1400 verschillende woorden herkennen. Het is ondoenlijk om voor zoveel woorden modellen op te slaan. Daarom worden tegenwoordig meestal fonemen als basiseenheden gebruikt.

Fonemen, of spraakklanken, zijn de bouwstenen van spraak. Voor ieder foneem is er een symbool (zie tabel 1). Bijvoorbeeld in het woord 'rekenwerk' zitten drie verschillende klanken 'e', die dus alle beschreven moeten worden met een ander foneemsymbool (resp. 'e:', '@' en 'E': 're:k@nwERk'). De 37 basiseenheden die in de spraakherkenner van OVIS gebruikt worden staan in tabel 1. Merk op dat er zowel voor de 'l' als de 'r' twee basiseenheden bestaan, namelijk voor en na een klinker. Dit is gedaan omdat een 'l' en 'r' voor en na een klinker vaak anders uitgesproken worden.

Tabel 1. De 37 gebruikte basiseenheden met per basiseenheid een voorbeeld van een woord waarin de klank voorkomt (de desbetreffende klank is vet en schuin weergegeven).

klinkers				medeklinkers					
i	<i>liep</i>	I	<i>lip</i>	p	<i>put</i>	n	<i>nat</i>	s	<i>sap</i>
e:	<i>leeg</i>	E	<i>leg</i>	b	<i>bad</i>	l	<i>lat</i>	z	<i>zat</i>
a:	<i>laat</i>	A	<i>lat</i>	t	<i>tak</i>	r	<i>rat</i>	S	<i>sjaal</i>
o:	<i>boom</i>	O	<i>bom</i>	d	<i>dak</i>	L	<i>bal</i>	j	<i>jas</i>
y	<i>buut</i>	Y	<i>put</i>	k	<i>kat</i>	R	<i>bar</i>	x	<i>licht</i>
2:	<i>deuk</i>	@	<i>gelijk</i>	N	<i>lang</i>	f	<i>fiets</i>	h	<i>had</i>
Ei	<i>wijs</i>	u	<i>boek</i>	m	<i>mat</i>	v	<i>vat</i>	w	<i>wat</i>
9y	<i>huis</i>	Au	<i>koud</i>						

Een spraakherkenner kan alleen woorden herkennen die in zijn woordenboek staan. Een stukje van het OVIS lexicon staat in tabel 2. In dit lexicon staan voor ieder woord twee vormen. Links de orthografische vorm, oftewel het woord zoals het geschreven wordt. Dit is een reeks van letters. En rechts het woord zoals het uitgesproken wordt. Dit is een reeks van foneemsymbolen, en wordt daarom een foneemtranscriptie genoemd. Voor iedere foneem is een akoestisch model aanwezig. Door fonemen te gebruiken als basiseenheden, kan volstaan worden met het opslaan van 35 modellen, een veel kleiner aantal dan het totaal aantal woorden.

Alvorens een spraakherkenner gebruikt kan worden, moet hij getraind worden. Hiervoor wordt een grote hoeveelheid spraak gebruikt, wat een corpus genoemd wordt. Het corpus bevat voor iedere uiting zowel het spraaksignaal als een beschrijving van de inhoud van die uiting (d.w.z. de woorden die uitgesproken zijn). Een dergelijke verzameling wordt een corpus genoemd. Tijdens training worden voor alle uitingen uit het corpus de woorden opgezocht in het lexicon en wordt ieder woord vervangen door zijn foneemtranscriptie. Zo worden foneemtranscripties voor de uitingen verkregen.

Tabel 2. Een gedeelte van het OVIS lexicon.

orthografie	foneemtranscriptie
ravenstein	ra:v@nstEin
rechtstreeks	rExstre:ks
rechtstreekse	rExstre:ks@
reis	rEis
reisgelegenheid	rEisx@le:x@nhEit
reizen	rEiz@
renesse	r@nEs@
rest	rEst
retour	r@tu~R
retourreis	r@tu~rEis
retourtje	r@tu~Rc@
reuver	r2:v@R
rheden	re:d@
rhenen	re:n@
richting	rIxtIN
ridderkerk	rId@RkERk

Vervolgens wordt de optimale oplijning tussen het spraaksignaal en de foneemtranscriptie automatisch bepaald met een algoritme dat het Viterbi-algoritme heet. Deze oplijning is in feite een segmentering, d.w.z. in het spraaksignaal worden de grenzen van ieder van de elementen uit de foneemtranscriptie bepaald. Voor de gevonden oplijning berekent het Viterbi-algoritme tevens een kans. Deze kans kan geïnterpreteerd worden als de kans dat het spraaksignaal en de foneemtranscriptie bij elkaar horen. Het Viterbi-algoritme berekent de oplijning met de grootste kans: de optimale oplijning.

In de trainingsfase worden het Viterbi-algoritme en een trainingslexicon gebruikt om een trainingscorpus geheel te segmenteren. Na segmentatie kan voor iedere basiseenheid (foneem) opgezocht worden welke stukken spraaksignaal uit het trainingscorpus hierbij horen. Voor ieder foneem worden alle bijbehorende stukken spraaksignaal statistisch verwerkt en wordt een stochastisch model berekend: een zgn. 'hidden Markov model'. Ieder aldus berekend akoestisch model is een statistische beschrijving van alle bij dat foneem behorende stukken spraaksignaal uit het trainingscorpus.

Daarnaast worden er taalmodellen getraind. De taalmodellen die in het huidige systeem gebruikt worden zijn een unigram (de kans op ieder woord) en een bigram (de kans op een sequentie van twee woorden). Deze taalmodellen worden getraind door in het corpus te tellen hoe vaak ieder woord respectievelijk iedere combinatie van twee woorden voorkomt. Deze waarden worden dan gedeeld door het totaal aantal woorden respectievelijk het totaal aantal woordparen.

Een spraakherkenner bestaat uit de akoestische modellen, de taalmodellen en het herkenningsexicon. In de herkenningfase wordt geprobeerd een onbekende uiting te herkennen. Dit gaat als volgt. De spraakherkenner genereert alle mogelijke sequenties van woorden. Omdat van te voren niet bekend is uit hoeveel woorden een uiting bestaat, is het aantal hypothesen gigantisch groot, zeker als de herkenner over een groot lexicon beschikt. Gelukkig worden alle hypothesen van begin af aan gescoord, dat wil zeggen dat bepaald wordt hoe waarschijnlijk iedere hypothese is gegeven het binnenkomende (onbekende) signaal. Hiervoor wordt weer het Viterbi-algoritme gebruikt, dat de optimale oplijning en de bijbehorende kans bepaalt. Het merendeel van de hypothesen blijkt dan

al na een paar stappen zoveel minder waarschijnlijk te zijn dan de favorieten, dat ze zonder enig gevaar geschrapt kunnen worden uit de lijst van mogelijke oplossingen. Op die manier blijven geheugenbeslag en rekentijd voor het scoren van de hypothesen binnen redelijke grenzen. Dit is belangrijk omdat een spraakherkenner in de praktijk 'real-time' moet werken. Uiteindelijk wordt de sequentie van woorden met de grootste kans gekozen, en dit is de herkende zin.

Applicaties

In de inleiding werden twee toepassingen genoemd, namelijk een 'handen-vrij telefoon' en een 'gesproken-dialoogsysteem'. Het wordt nu tijd om antwoord te geven op de vraag: "*Is dit toekomst of realiteit?*".

Bij een handen-vrij telefoon kies je het nummer met je stem. Dan heb je daarvoor zowel je handen als je ogen niet nodig, en dus kun je die bijvoorbeeld blijven gebruiken om de auto te besturen. Je kunt het nummer kiezen door de getallen uit te spreken. Desgewenst kunnen (veel gedraaide) nummers voorzien worden van een zelf gekozen gesproken label zoals bijvoorbeeld 'thuis' of 'James' (zie de inleiding). Dergelijke telefoons zijn inmiddels te koop.

Gesproken-dialoogsysteem worden inmiddels ook al gebruikt, met name voor het geven van informatie via de telefoon. Het hierboven beschreven OVIS is een voorbeeld van zo'n systeem in Nederland. Wat dat betreft is het realiteit. Deze programma's kun je echter nog niet in de PC-winkel om de hoek kopen. Daar moet nog even op gewacht worden.

Wat in dezelfde PC-winkel wel te koop is, zijn automatische spraakherkenners. Tegenwoordig kun je er al een kopen voor nog geen 100 gulden. Deze spraakherkenners kun je gebruiken om tekst te dicteren (spreken i.p.v. typen) of om je PC te besturen (je roept dan commando's). Als je dan nog een telefooninterface hebt, en je bent een handige knutselaar, dan zou het in principe mogelijk moeten zijn om via de telefoon bij u thuis dingen te besturen, zoals een verwarming of video. Spreekt dat u wel aan, probeer het dan eens uit. Een waarschuwing: bouw wel een beveiliging in. Automatische spraakherkenning is niet perfect; dus als verstaan is "Zet de verwarming op negentig graden" zou het fijn zijn als u duidelijk kunt maken dat het negentien moet zijn.

De auteur

dr Helmer Strik (40) studeerde natuurkunde aan de Katholieke Universiteit Nijmegen en promoveerde in 1994 aan dezelfde universiteit. Op de afdeling Taal & Spraak van deze universiteit verzorgt hij een aantal colleges in het kader van de opleidingen Taal, Spraak & Informatica (TSI) en Spraak- & Taalpathologie (STP). Sinds 1989 doet hij onderzoek naar automatische spraakherkenning. Verder doet hij onderzoek naar intonatie en het functioneren van de stembanden tijdens het spreken. Op dit moment is hij werkzaam als 'fellow' op een post-doctoraal project van de Koninklijke Nederlandse Akademie van Wetenschappen.
E-mail: strik@let.kun.nl.

Literatuur

Meer informatie over deze onderwerpen is te vinden in deze boeken:

- J.A. Markowitz. Using speech recognition. New Jersey: Prentice-Hall Inc., 1996.
- L.R. Rabiner & B.-H. Juang. Fundamentals of speech recognition. New Jersey: Prentice-Hall Inc., 1993.
- C. Schmandt. Voice communication with computers. New York: van Nostrand Reinhold, 1994.

Of op deze websites:

- <http://lands.let.kun.nl/TSPublic/strik/speech-sites.html>
Een lange lijst met links naar interessante 'speech sites'.
- <http://www.telecats.nl/artikel/>
Inleidend artikel op over spraaktechnologie.
- <http://www.ovr.nl/>
Site van OVR waar informatie over OVIS te vinden is.