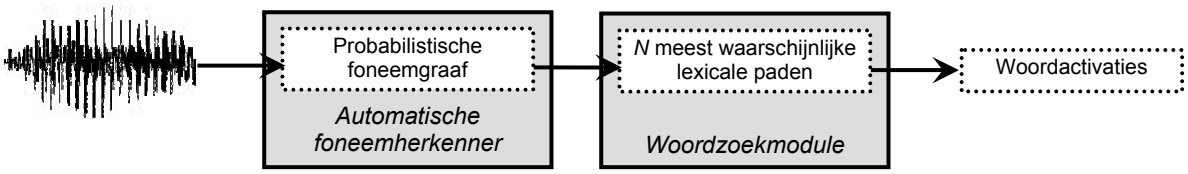
Wat is het stil aan de overkant…Onderzoek naar automatische en menselijke spraakherkenning

Wat vooraf ging

Op de middelbare school had ik een brede interesse. Ik vond talen leuk, maar ook de bètavakken lagen mij wel. Wat voor een vervolgstudie dan te kiezen met een eindexamenpakket van Nederlands, Engels, Frans, en geschiedenis, maar ook biologie, wiskunde A, scheikunde en natuurkunde? In de decanenkamer van mijn middelbare school vond ik informatie over de studie Taal, Spraak & Informatica aan de Faculteit der Letteren van de Katholieke Universiteit Nijmegen; een brede studie met alfa-, bèta- en gammaonderwerpen zoals grammatica, logica, programmeren, fonetiek, signaalverwerking en psycholinguïstiek. In 1995 ben ik met mijn studie begonnen. In maart 2000 ben ik afgestudeerd met een specialisatie in automatische spraakherkenning: het door een computer laten herkennen van door een mens geproduceerde spraak.

Mijn promotieonderzoek

Op 1 januari 2001 ben ik begonnen met mijn promotieonderzoek: psycholinguïstisch plausibele automatische spraakherkenning. Mijn promotieonderzoek lag tussen de twee wetenschapsgebieden van de menselijke spraakherkenning (MSH; een onderdeel van de psycholinguïstiek) en de automatische spraakherkenning (ASH). Het doel van het onderzoek naar menselijke spraakherkenning is het komen tot een volledig begrip van het menselijke spraakherkenningsproces. Om dit te bereiken worden laboratoriumexperimenten met luisteraars uitgevoerd. Op basis van de resultaten van deze experimenten worden theorieën over bepaalde aspecten van het menselijke spraakherkenningsproces opgesteld. Om deze theorieën vervolgens te testen worden computationele modellen gebouwd voor de simulatie van het menselijke spraakherkenningsproces. Bijna alle computationele modellen hadden dezelfde vereenvoudigende aanname, namelijk dat op de een of andere manier het spraaksignaal al omgezet is naar een reeks van klanksymbolen. Bestaande computationele modellen van menselijke spraakherkenning konden dus geen echte spraak herkennen. Deze tekortkoming maakt het lastig om alle aspecten van een theorie te testen. Automatische spraakherkenners, aan de andere kant, moeten wel in staat zijn om echte spraak te herkennen. Eén van de onderwerpen van mijn promotieonderzoek was dan ook om een computationeel model van menselijke spraakherkenning te bouwen dat in staat is echte spraak te herkennen. Een dergelijk computationeel model bestaat uit een aantal componenten. Figuur 1 geeft een overzicht van het door mij gebouwde computationele model SpeM.



Figuur 1. Een overzicht van de componenten van een computationeel model van menselijke spraakherkenning, dat echte spraak kan herkennen.

De eerste component, de automatische foneemherkenner, zet het spraaksignaal om in een probabilistische foneemgraaf. Een automatische foneemherkenner is een speciaal soort automatische spraakherkenner: in plaats van woorden herkend het klanken, oftewel, fonemen. De invoer van de automatische foneemherkenner bestaat uit het akoestische signaal. Met behulp van Bayes' Rule wordt bepaald welk woord (of woordreeks, of foneem) het meest waarschijnlijk is gegeven het akoestische signaal ($P(W|A)$):

$$P(W \mid A) = \frac{P(A \mid W) \bullet P(W)}{P(A)}$$

waarbij W het woord is (of de woordreeks, of het foneem in het geval van foneemherkenning); A is het akoestische signaal; $P(A|W)$ is de kans dat het akoestische signaal A ontstaat gegeven W. Deze kans volgt uit het akoestische model (in de vorm van een hidden Markov Model). Een akoestisch model (meestal voor ieder foneem één) is getraind op veel spraakdata en is een gemiddelde weergave van een bepaalde klank. $P(W)$ is de (a-priori) kans op W; $P(A)$ is de kans op het akoestische signaal A. Deze kans is moeilijk te schatten, maar voor het berekenen van de meest waarschijnlijke W gegeven A is deze kans niet vereist.

Tijdens het spraakherkenningsproces wordt het spraaksignaal eerst door een akoestische *pre-processor* gehaald waar featurevectoren geëxtraheerd worden uit het spraaksignaal. Vervolgens worden deze featurevectoren, die idealiter alle karakteristieken van het spraaksignaal beschrijven, vergeleken met de akoestische modellen (in het geval van woordherkenning worden de featurevectoren vergeleken met een opeenvolging van akoestische modellen die overeenkomen met woorden). Voor iedere featurevector wordt de match met ieder van de akoestische modellen berekend. Uiteindelijk wordt het foneem wiens akoestische model het beste matcht met de input herkend. De output van de foneemherkenner bestaat naast een sequentie van fonemen die het best matcht met de input ook uit een foneemgraaf. Een foneemgraaf bestaat uit één begin- en één eindknoop en bevat heel

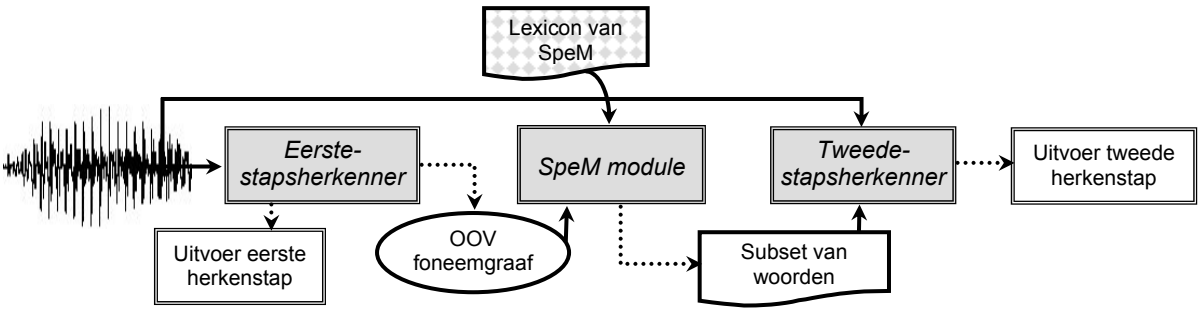
veel paden van het begin- naar het eindpunt. Ieder pad bestaat weer uit foneemreeksen. Een foneemgraaf is dus een zuinige representatie van het spraaksignaal.

De tweede component, de woordzoekmodule, zoekt vervolgens het beste (goedkoopste) pad door de zoekruimte gedefinieerd door de foneemgraaf en een lexicon (een woordenlijst afgebeeld in de vorm van een boom). De zoekmethode die gebruikt is gebaseerd op Viterbi; het is een tijdsynchrone breadth-first zoektocht. Deze module geeft vervolgens aan de uitvoer een lijst van de *N* meest waarschijnlijke paden bestaande uit de woorden uit het lexicon. In samenwerking met onderzoekers van het Max Planck Instituut (Nijmegen) en uit Cambridge, Groot-Brittannië, is deze module door mij geïmplementeerd. Hiervoor heb ik in 2002 twee maanden doorgebracht in Cambridge. In verschillende simulaties hebben we laten zien dat het door ons geïmplementeerde computationele model van menselijke spraakherkenning niet alleen daadwerkelijk spraak kan herkennen, maar ook belangrijke resultaten gevonden in experimenten met menselijke luisteraars kan simuleren.

Een tweede belangrijk aspect van mijn promotieonderzoek betrof de meerwaarde van SpeM voor automatische spraakherkenning. Automatische spraakherkenners kunnen alleen die woorden herkennen die aanwezig zijn in het lexicon. Woorden die niet aanwezig zijn in het lexicon, zogenaamde out-of-vocabulary woorden, oftewel OOVs, zullen op woorden gematcht worden die wel in het lexicon staan en zo dus voor een fout herkend woord zorgen. Binnen de automatische spraakherkenning wordt er veel onderzoek gedaan naar hoe dit probleem het beste opgelost kan worden. Het is namelijk onmogelijk om alle woorden in het lexicon van een automatische spraakherkenner op te nemen, aangezien er altijd weer nieuwe woorden ontstaan en/of gemaakt kunnen worden (dit geldt in het bijzonder voor talen zoals het Nederlands en Duits waar nieuwe woorden gevormd kunnen worden door twee zelfstandige naamwoorden aan elkaar te plakken, bijv. *huis* en *deur* wordt *huisdeur*).

In de zomer van 2004 ben ik voor 3 maanden naar de Computer Science and Artificial Intelligence Laboratory van het MIT in Cambridge, USA, geweest. Hier heb ik samen met onderzoekers van die groep onderzoek gedaan naar een mogelijkheid om het OOV-probleem op te lossen binnen het geautomatiseerde dialoogsysteem dat zij daar hebben draaien. Het MIT-dialoogsysteem geeft o.a. vluchtinformatie en de weersvoorspelling voor meerdere Amerikaanse steden. Gebruikers kunnen bellen met een telefoonnummer en krijgen vervolgens de gevraagde informatie van de computer. In dit onderzoek hebben we SpeM's mogelijkheid om foneemreeksen i.p.v. woorden te kunnen herkennen benut.

Figuur 2 geeft een overzicht van het systeem dat gebouwd is voor het onderzoek met MIT. In de eerste stap werd een spraakherkenner gebruikt die een relatief klein aantal woorden (<3K woorden) kon herkennen. Alle woorden die niet in zijn lexicon stonden werden als OOV aangemerkt en daarvoor werd een OOV foneemgraaf gecreëerd. Deze OOV foneemgraaf werd vervolgens aan SpeM als input gegeven; op basis van een veel groter lexicon (>50K woorden) zocht SpeM vervolgens de meest waarschijnlijke woorden gegeven de OOV foneemgraaf. De kracht van deze aanpak ligt er in dat SpeM ook 'onaffe' woorden zoals *amste* kan 'herkennen', waarna het onaffe woord uitgeschreven wordt naar een set van woorden die allemaal met de onaffe reeks fonemen beginnen. De hele set van meest waarschijnlijke woorden gedefinieerd door SpeM wordt vervolgens toegevoegd aan het lexicon van de tweede-stapsherkenner, waarna deze de laatste herkenstap doet. Dit onderzoek heeft laten zien dat het toevoegen van de woorden gevonden door SpeM aan het lexicon van de tweede-stapsherkenner, de herkenprestatie met zo'n 9.0% (absoluut) heeft verbeterd ten opzichte van de herkenprestatie van de eerste-stapsherkenner waar dus veel woorden ontbraken in het lexicon.



Figuur 2. Overzicht van het herkensysteem gebruikt voor het onderzoek met MIT.

Het vervolg

Inmiddels heb ik mijn proefschrift met succes verdedigd. Verder heb ik van NWO een Talentstipendium toegekend gekregen, die ik zal gebruiken voor een vervolgonderzoek aan de Speech and Hearing Research Group, Department of Computer Science, University of Sheffield, Groot-Brittannië.

Odette Scharenborg (O.Scharenborg@let.ru.nl)
Afdeling Taalwetenschap / Taal & Spraak