Tentamen Spraakherkenning en -synthese

David Weenink

30 maart 2007, zaal C.212, 14.00–17.00

Vermeld op iedere pagina je *naam*, je *studentnummer* en het *vervolgnummer per pagina*. Als je voor 16.30 u klaar bent, lever dan je tentamen in en verlaat rustig de zaal.

Het cijfer voor dit schriftelijk tentamen bepaalt je eindcijfer voor dit college. Dat eindcijfer wordt echter pas aan het OWI doorgegeven wanneer je ook alle 8 practicumverslagen hebt voltooid en ze zijn goedgekeurd.

Beantwoord onderstaande vragen zo nauwkeurig, compleet en gedetailleerd mogelijk, zonder in herhalingen te vervallen. De vragen zijn zo gesteld dat een kort antwoord volstaat. Een erg lang antwoord is bijna zeker een verspilling van tijd. Bij de beoordeling van dit tentamen wordt meer gelet op begrip dan op feitenkennis. Besteed niet te veel tijd aan afzonderlijke opgaven. Als een opgave te veel tijd vergt, probeer dan eerst een andere opgave.

1 Spreken en verstaan

1.a) Rubricering van klanken

Medeklinkers kunnen van elkaar worden onderscheiden door ze te rubriceren naar 'manier van articulatie', 'plaats van articulatie' en 'mate van stemhebbendheid'. Zo is een 'p' een stemloze labiale plofklank en 'z' een stemhebbende alveolaire fricatief.

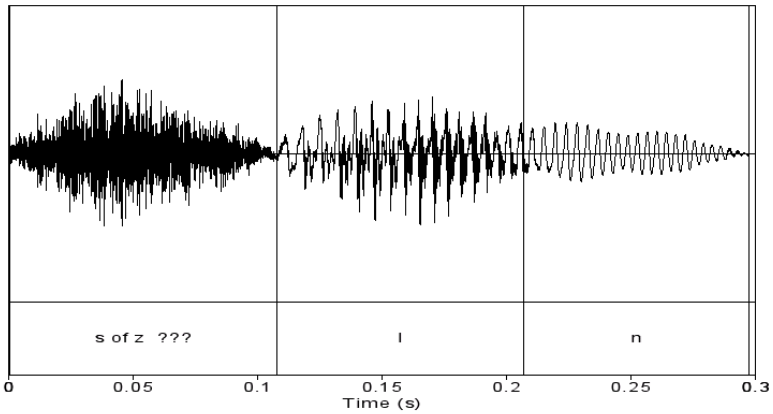
- Rubriceer 'b', 's', 't' en 'm'.
- Licht de bovengenoemde begrippen 'manier', 'plaats' en 'stem' nader toe.

1.b) Schrijven en spreken

Wat zijn de meest opvallende verschillen tussen geschreven tekst, zoals in 'dit is zichtbaar gemaakte spraak', en gesproken taal, zoals in /dIdIsIxbarx@makt@sprak/, onder andere in termen van

- de vluchtigheid van de informatie;
- de variabiliteit van de informatie, en
- het discreet of continu zijn van informatie?

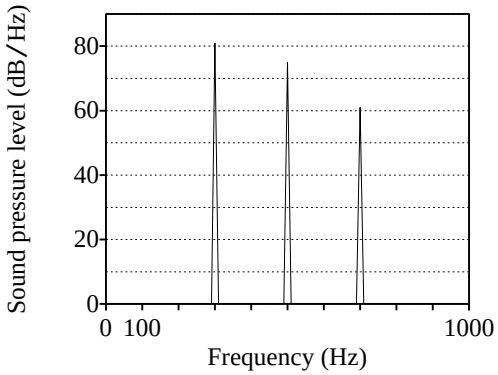
1.c) Productie



Figuur 1: Het laatste stukje van 'Is dit een vraagzin?'

Beschrijf van een drietal fonemen, namelijk de stemloze fricatief /s/, de korte ongeronde voorklinker /I/ en de dentale nasaal /n/ in het woorddeel 'zin' uit de uiting 'Is dit een vraagzin?', de wijze waarop ze door de mens worden geproduceerd (longen, stembanden, vorm mondkanaal, stand lippen, e.d.). Hiervoor zie je een oscillogram voor dit woorddeel plus de foneemgrenzen.

- Geef een indicatie van alle 3 de segmentduren in milliseconden.
- Teken ook een gestileerd spectrogram van dit woorddeel 'zin', incl. asbijschriften.
- Is uit bovenstaand oscillogram af te leiden of het eerste foneem als /s/ of als /z/ is uitgesproken?
- Vraagzinnen worden vaak uitgesproken met een stijgende intonatie aan het eind van de zin. Is dit hier ook het geval, en hoe kun je dat zien?



Figuur 2: Het spectrum van $s(t) = \cos(2\pi 300t) + 0.5\cos(2\pi 500t) + 0.1\cos(2\pi 700t)$

2 Het spectrum

In het amplitudespectrum $S(f)$ van een signaal $s(t)$ wordt de sterkte van een frequentiecomponent f_i weergegeven in dB/Hz (decibel per Hertz). Figuur 2 toont het spectrum van een signaal dat uit drie cosinuscomponenten bestaat met frequenties f_i van 300, 500 en 700 Hz (voor i gelijk aan 1, 2 en 3, respectievelijk). De amplitudes a_i van deze componenten zijn respectievelijk 1, 0.5 en 0.1. De functie die dit signaal in het tijddomein beschrijft is $s(t) = \cos(2\pi 300t) + 0.5\cos(2\pi 500t) + 0.1\cos(2\pi 700t) = \sum_{i=1}^{3} a_i \cos(2\pi f_i t)$. Als we met $S(f_i)$ de hoogte van een piek in het spectrum bij frequentie f_i aanduiden, dan is $S(f_i) = 20\log(a_i/a_r)$; a_r is een schaalfactor. In bovenstaande figuur is $S(300) = 81$ dB/Hz.
Als je weet dat $\log 2 \approx 0.3$, dan heb je voor de volgende onderdelen geen rekenmachientje nodig.

2.a) Component berekenen 1

Bereken $S(500)$ en $S(700)$.

2.b) Component berekenen 2

Bereken $S(300)$, $S(500)$ en $S(700)$ als we alle amplitudes in $s(t)$ twee keer zo groot maken.
$(s_2(t) = 2s(t) = 2\cos(2\pi 300t) + 1\cos(2\pi 500t) + 0.2\cos(2\pi 700t))$

2.c) Component berekenen 3

Bereken $S(300)$, $S(500)$ en $S(700)$ als we alle amplitudes in $s(t)$ twee keer zo klein maken.
$(s_3(t) = 0.5s(t) = 0.5\cos(2\pi 300t) + 0.25\cos(2\pi 500t) + 0.05\cos(2\pi 700t))$

2.d) Klinkerspectrum

Wat gebeurt er met de relative sterktes van de pieken in het spectrum van een klinker als deze alleen maar iets luider wordt uitgesproken?

3 Automatische Spraaksynthese

3.a) Spraaksynthesesystemen

Er zijn vier methoden voor spraaksynthese behandeld in het college:

- Articulatorische synthese
- Regel- of formantsynthese
- Diphone concatenation
- Unit selection

Geef de karakteristieken van ieder van deze methoden. Hou het zo kort mogelijk.

3.b) Tekst in, Spraak uit: Bewerkingsstappen

We onderscheiden ruwweg 6 stappen in de Tekst-naar-Spraak generatie.

- Tekstnormalisatie
- Letter/klank conversie
- Accentplaatsing
- Duurbepaling
- Intonatiegeneratie
- Spraaksynthese

Beschrijf het doel van ieder van deze stappen. Hou het zo kort mogelijk.

3.c) Eigenschappen van een goede stem

Voor het aanmaken van een nieuwe "stem" voor een difoon of unit selection TTS synthese moeten er nieuwe spraakopnamen gemaakt worden. Geef de eigenschappen waarop gelet moet worden bij de selectie van een goede spreker en bij het maken van de opname.

3.d) Evaluatie

Bij de evaluatie van automatische Text naar Spraaksynthese wordt vaak gebruik gemaakt van Semantically Unpredictable Sentences (bv, *het dorp kookte naar vogels*) en echte krantenzinnen. Vaak worden er ook hele paragrafen uit langere artikelen gebruikt.

Geef voor ieder van deze tekstcategorieën aan wat ermee getest wordt.

4 Automatische Spraakherkenning

4.a) Bayesian inference

$$\hat{W} = \underset{Words}{\operatorname{argmax}}\, P\,(\,Words|\,Observation) =$$

$$\underset{Words}{\operatorname{argmax}} \frac{P\,(\,Observation|\,Words) \cdot P\,(\,Words)}{P\,(\,Observation)}$$

Centraal in Automatische spraakherkenning staat de bovenstaande formule waarbij $\hat{W}$ de meest waarschijnlijke zin is, *Words* een kandidaat zin en *Observation* het waargenomen geluid. Leg uit wat de verschillende componenten van deze formule betekenen, en wat de hele formule beoogt.

4.b) Taalmodel

Leg uit wat een Finite-State en een N-gram taalmodel zijn. Wat zijn de verschillen?
Waarvoor worden taalmodellen gebruikt in de Automatische Spraakherkenning?
Hoe kun je deze taalmodellen trainen en wat voor soort corpora heb je daarvoor nodig?

4.c) Tweetraps decodering

Figuur 3: Tweetraps N-beste decoder.

In de figuur 3 zie je het schema van een tweetraps decodering. Hoe werkt dit wanneer je het vergelijkt met de standaard, eentraps, herkenner? Wat wordt bedoeld met de term "N-best"?
Wat zijn de extra mogelijkheden ten opzichte van een ééntraps herkenner?

4.d) Trainen van akoestische modellen

Bij het maken van een spraakherkenner voor in de auto worden de sprekers en de geluidstechnici met de auto de straat opgestuurd voor het maken van opnamen. Waarom? Waar kunnen ze het beste gaan rijden?

4.e) Akoestische features

Bij de spraakherkenning wordt intern niet de gewone golfvorm van het geluid of het spectrum gebruikt. Waarom niet? Wat voor een representatie wordt er wel gebruikt?

5 Dialoogsystemen

5.a) Grice's Maxims

Er zijn 4 "Maxims" van Grice die gebruikt worden om tijdens conversaties gevolgtrekkingen (*inferences*) te maken. In het Engels worden die aangeduid als de *Maxims of*: *Quantity*, *Quality*, *Relevance*, en *Manner*.
Leg uit wat met ieder van deze Maxims bedoeld wordt. Hou het zo kort mogelijk.

5.b) OVIS training

Bij de constructie van OVIS (Openbaar Vervoer Informatie Systeem) werd begonnen met een bestaand corpus van over de telefoon opgenomen spraak (het Polyphone corpus). Geef aan hoe de OVIS herkenner werd getraind voor zijn taak. Dat wil zeggen, beschrijf welk spraakmateriaal werd gebruikt, hoe men het verzamelde en hoe men voorbeelden van interacties verzamelde voor een taalmodel.

5.c) Beurtwisselingen

We kennen de volgende regels voor wie als volgende mag spreken in een conversatie (geciteerd uit het Engels):

- If during this turn the current speaker has selected A as the next speaker then A must speak next
- If the current speaker does not select the next speaker, *any* other speaker may take the next turn
- If no one else takes the next turn, the *current* speaker may take the next turn

Dialoogsystemen kunnen deze regels gebruiken om een conversatie te sturen door middel van *adjacency pairs*. Leg uit wat dit zijn en hoe ze gebruikt worden. Geef voorbeelden.

6 Voorbeelden van kleine spraaktechnologieprojecten

Tijdens het college zijn "drie" kleine studenten projecten behandeld: Een spraaksynthese voor het Fries, en twee cijferherkenners voor Kinyarwanda en Nederlands.

6.a) Motivatie

Technologie heeft ook altijd politieke, culturele, en socio-economische kanten. Voor het Fries en Kinyarwanda was nog geen spraaktechnologie beschikbaar. Bespreek in het kort de motivatie om spraaktechnologie voor deze twee, en vergelijkbare, talen te ontwikkelen. Gebruik daarbij de argumenten van de betreffende auteurs.

6.b) Friese TTS

De Friese spraaksynthese werkte met de Nederlandse difonen. Dit leverde spraak op met een niet zo beste kwaliteit. Beschrijf in het kort hoe je eenvoudig een Friese difoon set zou kunnen maken met Festival/Nextens. Negeer de uiteindelijke codering van de golfvorm (bijvoorbeeld met MBROLA). (Dit overlapt deels met de manier waarop spraakmateriaal verzameld is voor de twee cijferherkenners)

6.c) Evaluatie van Automatische Spraakherkenning

LEFT OUT SPEAKER	WORD RECOGNITION (%)	SENTENCE RECOGNITION (%)
Tom	99.57	85.71
Markus	99.78	72.60
Ork	99.43	89.13
Frans	99.78	81.63

LEFT OUT PERCENTAGE	WORD RECOGNITION (%)	SENTENCE RECOGNITION (%)
12	99.41	92.86
25	99.80	90.57
50	99.84	89.35

Boven: Testen op een nieuwe spreker, Onder: Testen op nieuwe zinnen

Hierboven ziet u de resultaten van een evaluatie van de Nederlandse cijferherkenner. In de bovenste tabel is telkens getest met spraakmateriaal van een spreker waarop niet getraind is. In de onderste tabel is telkens getest op zinnen waarop niet getraind is.

Leg in het kort uit hoe deze evaluatie in zijn werk gaat en wat het doel ervan is in relatie tot de bruikbaarheid van de herkenner.