Tentamen Spraakherkenning en -synthese

Rob van Son

21 december 2005

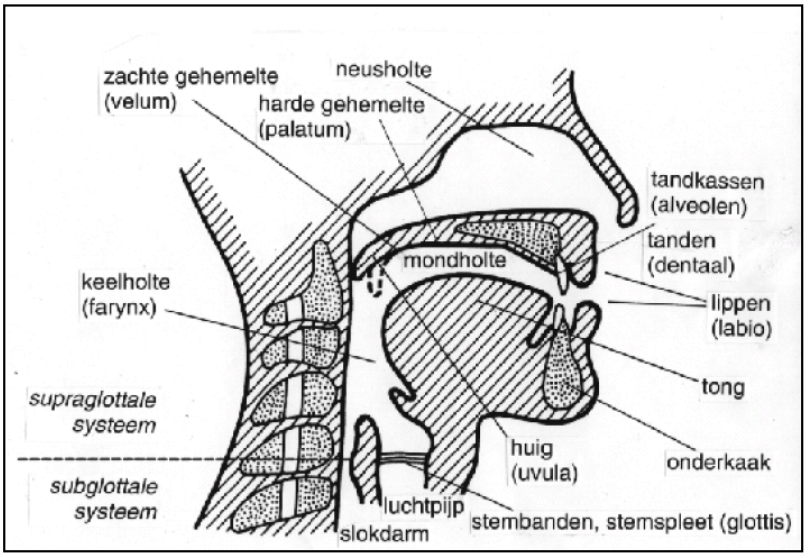
Vermeld op iedere pagina je naam, je studentnummer en het vervolgnummer per pagina. Als je voor 16.30 u klaar bent, lever dan je tentamen in bij de surveillant en verlaat rustig de zaal.

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. Nadere informatie hierover vind je op de Blackboard site.

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. Het kan geen kwaad relevante ervaringen vanuit het hoorcollege en de practicumopdrachten in je antwoorden te verwerken. 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.

Veel succes!

Rob van Son



Opgave **??**: Een dwarsdoorsnede van een menselijk hoofd

1 Spreken en verstaan

1.a) Spraakorganen

Geef aan de hand van de figuur aan welke organen van belang zijn voor het spreken en wat hun functie is. Hou het kort.

1.b) Grondtoon

Welk orgaan of welke organen zijn verantwoordelijk voor de hoogte van de F_0? Hoe verandert dit orgaan of veranderen deze organen de F_0? Wat zijn de verschillen tussen de stemmen van mannen, vrouwen en kinderen?

1.c) Klinkers

Beschrijf aan de hand van de bovenstaande afbeelding hoe een klinker wordt gevormd.

1.d) Plosieven & Fricatieven

Beschrijf aan de hand van de bovenstaande afbeelding hoe een plosief en een fricatief worden gevormd.

1.e) Articulatie

Welke bewegingen zijn nodig om het woord “lam” (/lɑm/) uit te spreken?

2 Automatische Spraaksynthese

2.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.

2.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.

2.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.

2.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.

2.e) Blizzard Challenge

De "Blizzard Challenge 2005" heeft tot doel om betere TTS technologieën te vinden door een internationale "wedstrijd" tussen TTS systemen te organiseren. Voor dit doel kregen alle deelnemers 4 centraal aangemaakte spraak corpora met 1200 fonetisch gebalanceerde zinnen om een nieuw TTS systeem te maken. De zinnen waren geselecteerd uit Project Gutenberg. Daarna kregen de deelnemers testzinnen toegestuurd om te genereren.

Leg in het kort uit waarom deze opzet is gekozen en niet aan iedere deelnemer gevraagd is om de test-zinnen met hun bestaande systeem te synthetiseren.

3 Automatische Spraakherkenning

3.a) Bayesian inference

$$\hat{W} = \underset{Words}{argmax}\ P\left(Words|Observation\right) = \underset{Words}{argmax} \frac{P\left(Observation|Words\right) \cdot P\left(Words\right)}{P\left(Observation\right)}$$

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.

3.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?

3.c) Tweetraps decodering

Opgave **??**: Tweetraps N-beste decoder

In de figuur Opgave **??** 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?

3.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?

3.e) Akoestische features

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

4 Dialoogsystemen

4.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.

4.b) Wat is een *practical dialog*?

Leg uit wat een *practical dialog* is en geef een paar voorbeelden. Waarom is dit onderscheid van belang voor de spraaktechnologie?

4.c) 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.

4.d) 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.

4.e) Grounding

Een belangrijk principe in de structuur van conversaties is de zogenaamde *Grounding*. Tijdens het college en in de literatuur zijn verschillende voorbeelden hiervan behandeld. Leg uit wat er met *Grounding* bedoeld wordt. Geef enkele voorbeelden hoe een (geavanceerd) dialoogsysteem dit zal gebruiken.

5 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.

5.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. Geef ook aan waar de, socio-economische, verschillen tussen deze twee talen liggen.

5.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)

5.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 of 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.