

Hertentamen Spraakherkenning en -synthese

Rob van Son

12 juni 2008

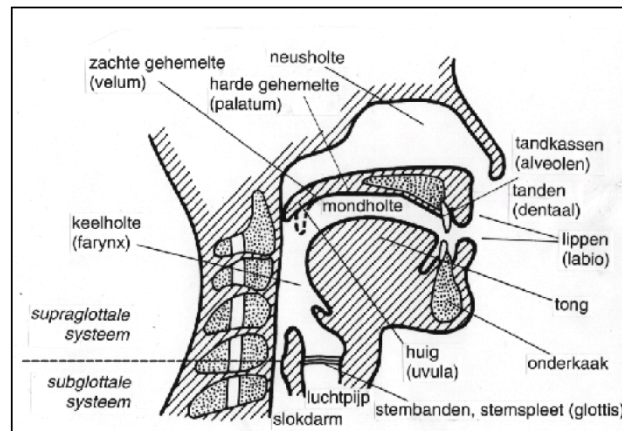
Vermeld op iedere pagina je naam, je studentnummer en het vervolgnummer per pagina. Als je voor 13.00 u klaar bent, lever dan je tentamen in bij de surveillant en verlaat rustig de zaal.

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 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 1: Een dwarsdoorsnede van een menselijk hoofd

1 Spreken en verstaan

1.a) Spraakorganen

Medeklinkers worden gemaakt door een constrictie in het spraakkanaal te maken. Leg aan de hand van de figuur aan waar en hoe deze constricties gemaakt worden voor de verschillende fricatieven en plosieven. Wat is het verschil tussen deze twee klassen?

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) Tweeklanken

Beschrijf aan de hand van de bovenstaande afbeelding hoe een diphthong, zoals een /ei/, wordt gevormd.

1.d) Reductie

Wat zijn de meest opvallende verschillen tussen geschreven tekst, zoals “natuurlijk was ik er”, en gesproken taal /tyləkwazəkər/, onder andere in termen van

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

Licht ook toe welke additionele informatie (b.v. over akoestiek, emotie, e.d.) in het spraaksignaal zit, die in geschreven tekst ontbreekt.

2 Automatische Spraaksynthese

2.a) Spraaksynthesesystemen

Er zijn vier methoden voor spraaksynthese behandeld in het college:

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

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

2.b) Text-to-Speech synthesis

Beschrijf in het kort wat er zoal nodig is om een gedrukte-tekst-op-papier om te zetten in verstaanbare synthetische spraak voor ieder van de onderstaande taken:

- text preprocessing;
- grafeem-foneemconversie via uitspraaklexicon en/of regels
- prosodiecontrole (spreektempo, woordklemtoon, zinsaccent, intonatie, frasering)
- feitelijke signaalgeneratie.
- Geef tenslotte aan hoe de natuurlijkheid en de verstaanbaarheid van een spraaksynthesesysteem kan worden geëvalueerd? Geef tenminste twee voorbeelden van mogelijk stimulusmateriaal om zo'n evaluatie uit te voeren.

2.c) Prosodie

Het zinsaccent wordt eerst en vooral gerealiseerd via toonhoogtebewegingen op de beklemtoonde lettergrepen.

- Geef van onderstaande zin aan welke lettergrepen beklemtoond zijn, en welk(e) woord (en) jij zinsaccent zou geven.

In welke zaal wordt het tentamen afgenomen?

- Teken vervolgens een gestileerde toonhoogtecontour die bij deze classificaties past.

- Geef tevens van iedere klinker aan of de duur korter, langer, dan wel ongeveer gelijk aan de gemiddelde duur van die klinker zal zijn, en geef aan waarom je dat denkt.

3 Automatische Spraakherkenning

Bekijk de taak om een spraakherkenner te bouwen voor een automatische informatiekiosk in een stationshal. De herkenner moet lopende spraak van willekeurige gebruikers herkennen. De herkende tekst wordt gebruikt voor het opzoeken van informatie uit een taakgerichte database. Beschrijf heel in het kort de volgende aspecten:

3.a) Training en trainingsmateriaal voor het akoestische gedeelte

Verzameling en selectie van opnamen voor HMM foneemmodellen

3.b) Structuur van het taalmodel in relatie tot de taak

Bigram of anderszins

3.c) Trainingsmateriaal voor het taalmodel

Verzamelen en structuur.

3.d) Evaluatie van de herkenner

Hoe wordt de “kwaliteit” van de herkenner geëvalueerd in relatie tot de beoogde taak.

4 Word Error Rate

De performance van een spraakherkenner wordt vaak uitgedrukt in word error rate (WER), zijnde de som van het aantal substituties (S), inserties (I) en deleties (D), gedeeld door het totaal aantal woorden in de referentie (W), maal 100

$$WER = \frac{S + I + D}{W} \cdot 100$$

Neem iedere letter als een woord. Als de woordsequentie van de ingesproken letterreeks was: *NMLKJIHGF*, wat is dan, uitgaande van een optimale match:

4.a) De WER bij een herkende output: *NKMKJIIGFN*

4.b) De WER bij een herkende output: *LKJJIH?*

Licht je antwoorden toe en geef aan waar er Substituties, Inserties en Deleties hebben plaatsgevonden.

5 HMM ASR

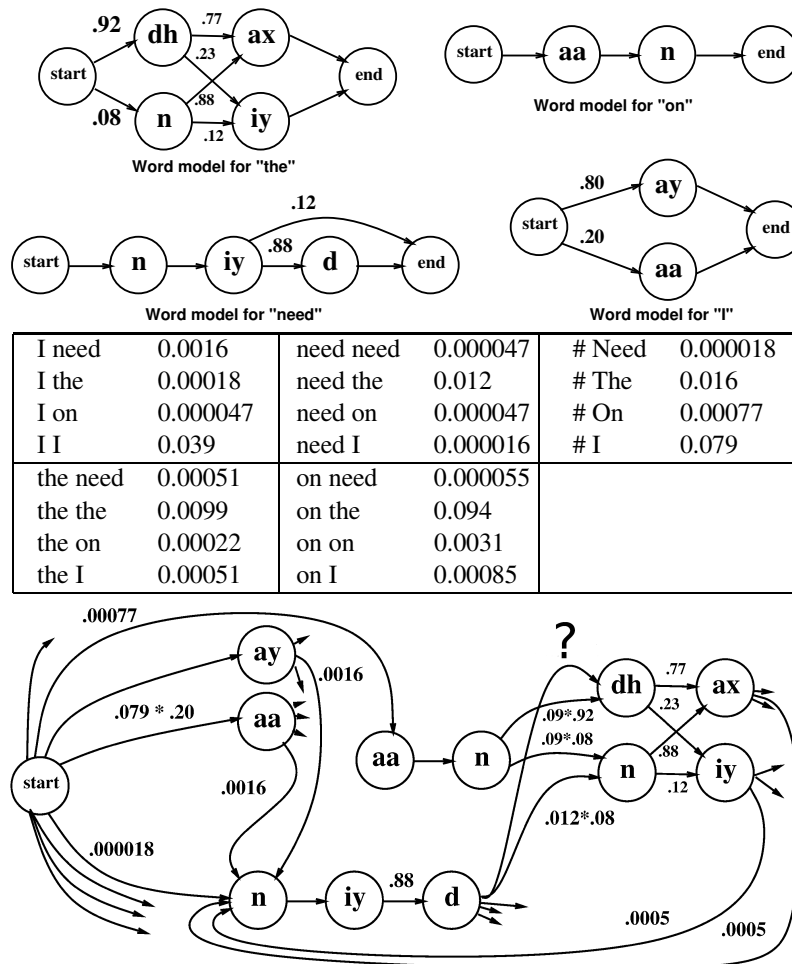


Figure 1: Hier zijn uit het boek van Jurafsky & Martin (2000) allereerst 4 woordmodellen voor 4 Amerikaanse woorden (*the*; *on*; *need*; *I*) weergegeven. In de daarna weergegeven tabel zijn de bigramwaarschijnlijkheden voor een aantal woordcombinaties weergegeven. Tenslotte is een stukje van een min of meer volledig HMM-herkenningsmodel weergegeven.

- Beschrijf hoe je dit model moet interpreteren.
- Geef tevens aan wat de betekenis is van de getallen $0.079 * 0.20$ en van het getal 0.23 in het woordmodel voor *the*.
- Geef tenslotte aan welk getal of getallencombinatie moet staan bij de met een vraagteken gemarkeerde pijl in de onderste figuur.

6 Dialoogsystemen

In het MATIS project vragen gebruikers informatie over treinreizen op door een formulier op het computerscherm in te vullen met, naar keuze, spraak, muisklikken of door tekst in te typen. Gebruikers kunnen voor ieder veld of keuze-mogelijkheid in het formulier zelf bepalen hoe ze het willen invullen. MATIS geeft naast spraakuitvoer ook altijd de al ingevulde velden van het formulier op het scherm weer. Het spraakherkennings en -synthese gedeelte van MATIS is gelijk aan dat van het OVIS project (uit: A spoken dialogue system for public transport information).

6.a) Trainingscorpora

MATIS werkt met een microfoon aan de computer, OVIS werkte over de telefoon zonder beeldscherm. Geef aan waarin het trainingsmateriaal voor de MATIS spraakherkenner zal verschillen van dat van de OVIS spraakherkenner. Zeg zowel iets over het tekstcorpus als over het spraakcorpus.

6.b) Dialoogmanagement

De spraak invoer van Matis kan door de gebruiker aan- en uitgezet worden door op een knop op het formulier te klikken. Een belangrijk aspect van MATIS is dat gebruikers het systeem op willekeurige momenten kunnen onderbreken met een muisklik. Hoe zou het OVIS dialoogmanagement systeem veranderd kunnen worden om dit mogelijk te maken?