

Deeltentamen Spraakherkenning en -synthese

Rob van Son

14-17 uur, 20 oktober 2008
REC-A AB.42 (Zaal B)

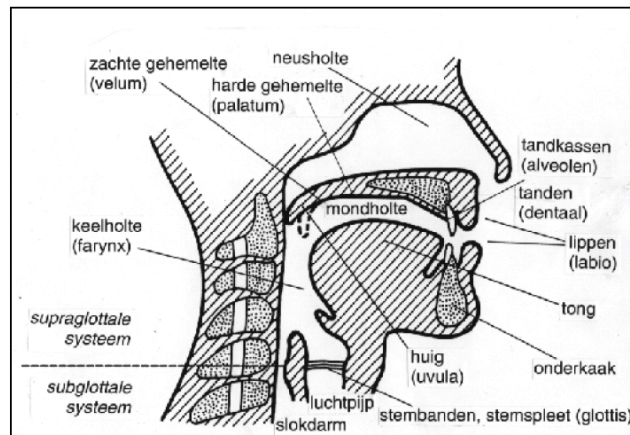
Vermeld op iedere pagina je naam, je studentnummer en het volgnummer per pagina. Gebruik voor elke opgave (1-4) een apart vel. Als je voor 17.00 uur klaar bent, lever dan je tentamen in bij de surveillant en verlaat rustig de zaal.

Het cijfer voor dit schriftelijk tentamen bepaalt de helft van je eindcijfer voor dit college. Dat *eind*cijfer wordt echter pas aan het OWI doorgegeven wanneer je ook alle practicumverslagen voor dit onderdeel 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 ter 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



Figuur 1: Een dwarsdoorsnede van een menselijk hoofd

1 Spreken en verstaan

1.a) Fonemen

Hoeveel fonemen heeft het woord *schooltijd*? Schrijf ze op.
Rubriceer deze fonemen op manier en plaats van articulatie.

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) Articulatie van het woord *schooltijd*

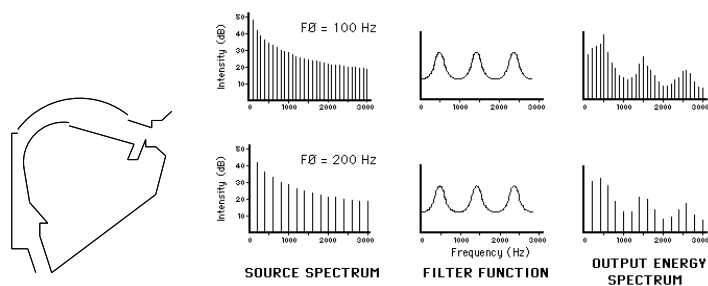
- i Geef van ieder klank aan of er een constrictie in het spraakkanaal gemaakt wordt. Zo ja, geef dan aan de hand van de figuur aan, waar die constrictie gemaakt wordt en met welke “organen” dit gebeurt.
- ii Beschrijf aan de hand van de bovenstaande figuur hoe de articulatie van de eerste twee klanken van het woord verloopt. Eindig met de overgang naar de derde klank.

1.d) Intonatie

Een declaratieve (of bevelende) zin eindigt in het Nederlands meestal met een dalende intonatie, een ‘L%’ in ToBI/ToDI. Met PSOLA (*Pitch Synchronous Overlap Add*), of de daarvan afgeleide MBROLA, techniek kun je de intonatie van spraak aanpassen in de golfvorm. Hoe zou je de pitch perioden in de golfvorm van het gesproken woord *schooltijd!* met een vlakke intonatie moeten aanpassen zodat het woord eindigt met een dalende intonatie?

Bv, in het zinnetje: *Doe dat binnen schooltijd!*

Let op: Het is niet de bedoeling dat je uitlegt hoe PSOLA werkt, maar alleen hoe je de plaats van de pitch perioden aanpast!



Figuur 2: Het bron filter model

2 Spraakklanken, spraakanalyse en corpora

2.a) Het bron filter model

Hierboven staat een illustratie van het bron-filter model. Leg aan de hand van bovenstaande figuren uit wat bedoeld wordt met formanten en F_0 . Hou het kort. Waarom zijn de formanten van mannenstemmen gemakkelijker te meten dan die van vrouwen of kinderen?

2.b) Formanten

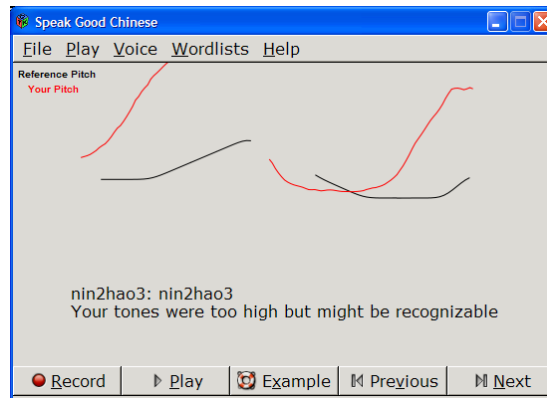
F_0 en formanten worden tijdens het spreken bepaald door verschillende spraakorganen. Geef aan welke spraakorganen een rol spelen bij het vormen van de F_0 en formanten van de a in het woordje *kat*.

2.c) Corpora

Het Corpus Gesproken Nederlands bevat ongeveer 8 miljoen gesproken woorden in verschillende spreekstijlen. In de beschrijving van het ontwerp van het CGN staat het volgende stukje tekst:

De omvang van de component 'voorgelezen tekst' heeft enige toelichting. Bij de start van het project was voorzien dat het corpus tien miljoen woorden min of meer spontaan gesproken Nederlands zou omvatten. Vanwege de aard van de in deze component aangegevoerde data - het betreft hier voorgelezen spraak die haar oorsprong vindt in geschreven teksten die bedoeld zijn gelezen te worden - was er weinig reden dit materiaal in het corpus op te nemen. Onder druk van de grote behoefte die er onder spraaktechnologen bestaat aan dergelijke data werd echter besloten toch een component van die aard op te nemen.

- i Leg uit waarom spraaktechnologen behoefte hebben aan voorgelezen spraak? Concentreer je daarbij op de behoefte van Tekst-naar-Spraak (TTS) technologieën.
- ii Voor de ontwikkeling van TTS synthese wordt ook gebruik gemaakt van spontane interacties tussen mensen en mens en computer. Waarvoor kunnen opnamen in deze "stijl" gebruikt worden?



Figuur 3 Het SpeakGoodChinese programma

3 SpeakGoodChinese

Het programma SpeakGoodChinese helpt studenten met het oefenen van de uitspraak van het Mandarijn Chinees

3.a) Toontalen

Mandarijn is een toontaal. Wat is een toontaal en hoe verschilt een toontaal van het Nederlands, Engels of Duits?

Toontalen geven problemen in het taalonderwijs. Geef enkele van die problemen en leg uit hoe een applicatie als SpeakGoodChinese docenten en studenten kan helpen bij het oplossen ervan. Wat is het belangrijkste educatieve aspect van het gebruikers interface?

3.b) Tone Synthese

SpeakGoodChinese synthetiseert tonen aan de hand van een model. De tonen die gesynthetiseerd worden zijn “extreem” en klinken over-gearticuleerd en nogal onnatuurlijk. Wat zou het effect hiervan op de studenten zijn? Zouden de docenten het nodig vinden dit te veranderen? Waarom wel of niet?

3.c) Spraaksynthese

Stel je zou een unit-synthese willen maken zodat SpeakGoodChinese alle één- en tweelettergrepige woorden kan laten horen als “normale” spraak.

Wat voor spraakmateriaal zou je hiervoor nodig hebben?

3.d) Evaluatie

Beschrijf in het kort hoe je de effectiviteit van een applicatie als SpeakGoodChinese zou kunnen evalueren?

4 Automatische spraaksynthese

4.a) Spraaksynthesesystemen

Er zijn vier methoden voor spraaksynthese behandeld in het college:

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

Welke methode zou het meest geschikt zijn voor een goedkope synthese voor *eenvoudige* uitingen, bv, *twéé-lettergrepinge* woorden in SpeakGoodChinese?

Unit-selectie en difoon-synthese lijken op elkaar, maar leveren duidelijk een verschillende kwaliteit van de synthese. Bij het maken van een spraaksynthese voor een nieuwe taal waar nog weinig of geen bestaande corpora en TTS systemen zijn zal een afweging gemaakt moeten worden tussen kosten en kwaliteit. Wat zijn de belangrijkste aspecten in een kosten-baten analyse tussen unit-selectie en difoon synthese?

4.b) Tekst in, Spraak uit: Bewerkingsstappen

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

- | | |
|-----------------------------------|-----------------------|
| I. Tekstnormalisatie | IV. Duurbepaling |
| II. Letter/klank conversie | V. Intonatiegeneratie |
| III. Accentplaatsing en frasering | VI. Spraaksynthese |

Beschrijf in het kort het verband tussen de eerste twee stappen, *Tekstnormalisatie* en *letter/klank conversie*. Gebruik voorbeelden uit de tekst van deze opgave.

4.c) Letter/klank conversie

De meest efficiënte manier om de letter/klank conversie in het Nederlands of Engels te doen is om een simpel uitspraakwoordenboek te gebruiken. In talen zoals het Duits, Fins, of Turks gaat dit mis omdat ze zoveel samengestelde woorden bevatten. Beschrijf (héél kort) een methode om deze conversie aan te passen voor samengestelde woorden in het Duits en Nederlands.

4.d) Prosodic Phrasing

Fraseren is het opdelen van lange uitingen in “uitspreekbare” brokken. Daarbij wordt gekeken naar interpunctie (b.v., komma’s). Welke mogelijkheden zijn er om een zin in een tekst op te breken in prosodische frasen als de interpunctie ontbreekt? Wat zijn hierbij de problemen?

```

<speak version="1.0" xml:lang="en-GB">
  Hello, how are you?
  <prosody rate="x-fast" >
    This sentence is spoken fast
  </prosody>
  <prosody pitch="x-low">
    This sentence is spoken low pitch
  </prosody>
  <prosody pitch="medium">
    This sentence is spoken medium pitch
  </prosody>
  <prosody pitch="x-high">
    This sentence is spoken high pitch
  </prosody>
  <prosody rate="fast">
    This sentence is spoken fast
  </prosody>
  <emphasis level="strong">
    This sentence is spoken with stress
  </emphasis>
</speak>

```

Figuur 4.e) Voorbeeld van Speech Synthesis Markup Language (SSML)

4.e) SSML

Speech Synthesis Markup Language (SSML) is een XML “taal” om TTS systemen te vertellen hoe woorden en zinnen uitgesproken behoren te worden. In principe kan met SSML elk aspect van de synthese, van spreker en tempo tot individuele fonemen, direct aangestuurd worden. Het wordt meestal gebruikt om de Tekst-naar-Prosodie omzetting aan te passen.

SSML is vooral belangrijk als de computer de tekst zelf aanmaakt. Het programma weet dan namelijk vaak waar de tekst over gaat en wat de gebruiker ermee wil. In het programma kan dan bepaald worden hoe de zinnen, frasen, en woorden het best uitgesproken kunnen worden om het beoogde doel te bereiken.

Neem als voorbeeld een programma dat een lokaal weerbericht aanmaakt op grond van gegevens over regenval, temperatuur, wind, en bewolking.

Geef heel in het kort aan welke specifieke aspecten van de Tekst-naar-Spraak omzetting met SSML markup verbeterd kunnen worden? Gebruik voorbeeldzinnen om je beschrijving te illustreren.

Het gaat hierbij om het principe. Het is voor deze opgave niet nodig om te weten welke tags er in SSML zitten en wat ze precies doen.