

Deeltentamen Spraakherkenning en -synthese

Rob van Son

9-12 uur, 25 oktober 2007
GEBA/A4.04(HC)

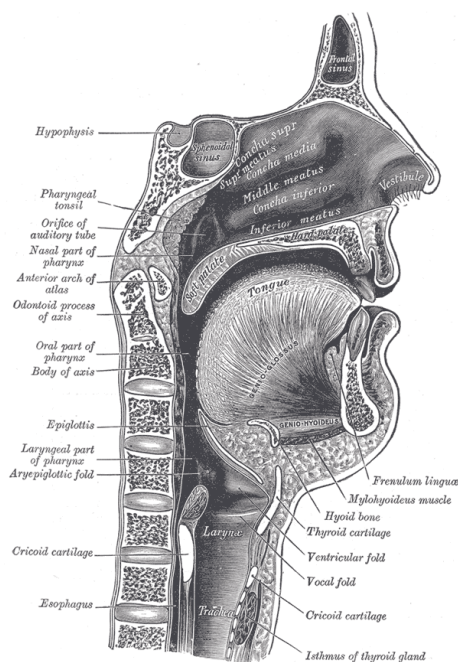
Vermeld op iedere pagina je naam, je studentnummer en het volgnummer per pagina. Gebruik voor elke opgave (1-4) een apart vel. Als je voor 12.00 uur klaar bent, lever dan je tentamen in bij de surveillant en verlaat rustig de zaal.

Het cijfer voor dit schriftelijk tentamen bepaalt de helft van je eindcijfer voor dit college. Dat eindcijfer 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 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



Figuur 1: Een dwarsdoorsnede van een menselijk hoofd

1 Spreken en verstaan

1.a) Fonemen

Hoeveel fonemen heeft het woord *viering*? Schrijf ze op.

1.b) Rubriceren van klanken

Rubriceer de fonemen van *viering* op manier en plaats van articulatie.

1.c) Articulatie 1

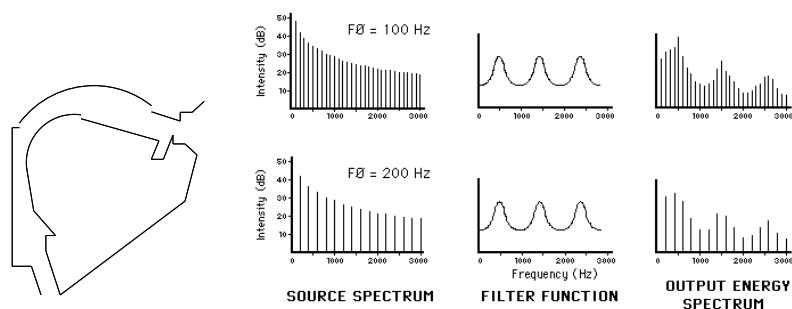
Geef van ieder klank in *viering* aan of er een constrictie in het spraakkanaal gemaakt wordt. Zo ja, geef dan aan de hand van de figuur aan, waar die constrictie gemaakt wordt en met welke “organen” dit gebeurt.

1.d) Articulatie 2

Beschrijf aan de hand van de bovenstaande figuur hoe de articulatie van de laatste klank in het woord *viering* verloopt, uitgaande van de voorlaatste klank.

1.e) Intonatie

Een vraagzin eindigt in het Nederlands meestal met een stijgende intonatie. Hoe zou je aan de golfvorm van het gesproken woord *viering?* kunnen zien dat dit woord vragend is uitgesproken?



Figuur 2: Het bron filter model

2 Spraakklanken, spraakanalyse en corpora

2.a) Het bron filter model

Leg uit wat bedoeld wordt met het bron-filter model aan de hand van bovenstaande figuren. Hou het kort.

2.b) Formanten

Hoe worden formanten gevormd tijdens het spreken en wat is hun rol in de spraak? Verwijs in je uitleg naar de dwarsdoorsnee van het hoofd en de figuren aan de rechterkant.

2.c) Sprekerverschillen

Waarom zijn de formanten van mannenstemmen gemakkelijker te meten dan die van vrouwen of kinderen?

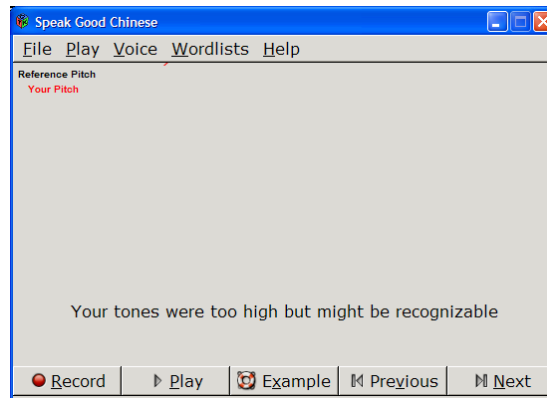
2.d) Transcriptie

Het Corpus Gesproken Nederlands bevat ongeveer 1 miljoen woorden waarvan de automatische fonetische transcriptie en de woordgrenzen handmatig gecontroleerd zijn. Er is ook een automatische oplijning van de fonemen gemaakt. Bij de aanmaak van het Corpus is in een vroeg stadium besloten om de oplijning van de fonemen niet handmatig te controleren.

Beargumenteer waarom het handmatig controleren van foneemgrenzen zo arbeidsintensief is. Betrek daarbij de betrouwbaarheid van de oplijning tussen de uitvoerders.

2.e) Corpora

Het Corpus Gesproken Nederlands bevat ongeveer 8 miljoen gesproken woorden in verschillende spreekstijlen. Er zijn *tekst* corpora van meer dan een miljard woorden verdeeld over verschillende tekststijlen. Geef de redenen waarom verschillende spreek- en tekststijlen gebruikt worden. Geef voorbeelden van spreek- en tekststijlen.



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?

Waarom hebben Nederlandse studenten zoveel moeite om Chinees te leren?

3.b) Tonen

Beschrijf in het kort de tonen van het Mandarijn Chinees. Gebruik zo mogelijk figuren.

3.c) Dynamic Time Warping

SpeakGoodChinese gebruikt Dynamic Time Warping (DTW) om de tonen te herkennen. Beschrijf in het kort hoe DTW werkt en hoe SpeakGoodChinese deze techniek gebruikt.

3.d) Spraaksynthese

SpeakGoodChinese heeft de beschikking over een verzameling van audio-opnamen van alle 1200 verschillende lettergrepen van het Mandarijn Chinees, inclusief alle voorkomende tonen. Alle woorden zijn door dezelfde spreker op een consistente manier ingesproken.

Stel je zou een difoonsynthese willen maken zodat je ook meerlettergrepige woorden kunt laten horen. Waarom kun je deze audiobestanden daar niet voor gebruiken?

Een alternatieve manier van synthese zou zijn om de audiofiles van de afzonderlijke lettergrepen achter elkaar te plakken. Wat voor problemen kun je hierbij verwachten?

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 *embedded applications*, bv, speelgoed?

Welke methode levert de beste kwaliteit?

Stel er moet een Tekst-naar-Spraaksynthese systeem gemaakt worden voor een nieuwe taal met weinig of geen digitale materialen en weinig middelen, bv, *Kinyarwanda* in Rwanda. Welke van deze methoden is dan het meest geschikt?

Beargumenteer in het kort je keuzes.

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 | VI. Spraaksynthese |

Beschrijf in het kort de eerste stap, *Tekstnormalisatie*, aan de hand van de tekst van deze opgave.

Wat is de meest efficiënte manier om de letter/klank conversie in het Nederlands te doen?

4.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.

4.d) Evaluatie

Bij de evaluatie van automatische Text naar Spraaksynthese wordt vaak gebruik gemaakt van zinnen met ambigue woorden (bv, *Now we will say cold/colt again.*) en fragmenten uit eerder uitgeschreven (mens-machine) dialogen (bv *Yeah, I guess it will and something downtown please.*).

Geef voor deze tekstcategorieën aan wat ermee getest wordt.

4.e) Blizzard Challenge

De “Blizzard Challenge” heeft tot doel om betere TTS technologieën te vinden door een internationale “wedstrijd” tussen TTS systemen te organiseren. Voor dit doel kregen alle deelnemers in 2005, 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 zelf een stem aan te maken en zelf zinnen uit te zoeken om hun systeem te testen.