

Tentamen Spraakherkenning en -synthese

Rob van Son

25 maart 2008

Vermeld op iedere pagina je naam, je studentnummer en het volgnummer per pagina. Gebruik voor elke opgave (1-5) 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 her-tentamen bepaalt je eindcijfer voor dit college. Dat eindcijfer wordt echter pas aan het OWI doorgegeven wanneer je ook alle practicumverslagen voor dit college 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. Reken op **9 minuten gemiddeld per sub-onderdeel** als je het antwoord weet. 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

1 Spreken en verstaan

1.a) Medeklinkers

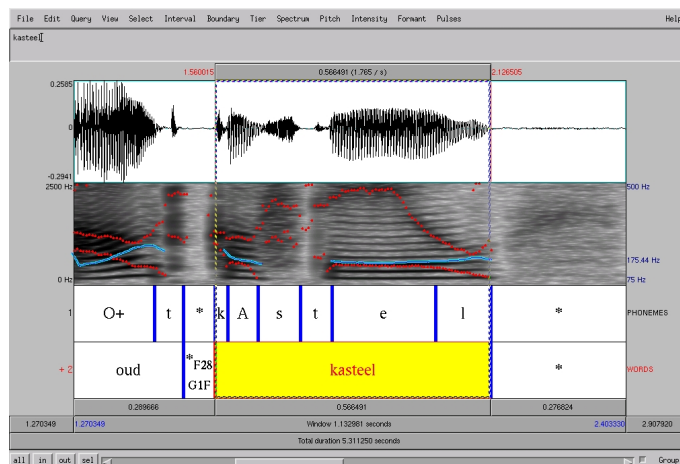
Medeklinkers worden ingedeeld in klassen gebaseerd op de *plaats van articulatie*. Beschrijf van iedere klasse hieronder de plaats van articulatie in het Nederlands:

- labiale klanken, zoals de /b/
- dentale klanken, zoals de /d/
- alveolaire klanken, zoals de /s/
- palatale klanken, zoals de /j/
- velaire klanken, zoals de /k/
- uvulaire klanken, zoals de /g/
- glottale klanken, zoals de /h/

1.b) Reductie

Leg uit waarom het woorden *Koninklijke Marine* vaak uitgesproken worden als /koːləkəmrinə/? Wat er gebeurt met de klinkers in zo'n geval?

1.c) Het bron-filter model



Figuur 1: De golfvorm, het spectrogram, de F_0 en de formanten van het woord *kasteel*

Beschrijf aan de hand van de fonemen van het woord *kasteel* in het plaatje hierboven wat de geluidsbron is bij ieder foneem en hoe het geluid van de bron beïnvloed wordt door de vorm van het spraakkanaal. Geef aan hoe je dit terugziet in het bovenstaande plaatje.

1.d) Spraakcorpora

Spraakcorpora bevatten meta-data. Wat wordt daarmee bedoeld? Geef voorbeelden van spraak specifieke meta-data.

Alle data in taal- en spraakcorpora moeten “genormaliseerd” worden. Geef aan wat normalisatie betekent. Gebruik een spraakcorpus voor het trainen van een spraakherkenner voor gebruik in auto's als voorbeeld.

2 Spraaksynthese

2.a) Prosody

- Wat is de functie van de F_0 in de zin?
- Wat zijn de effecten van klemtoon en accent op de fonemen in een beklemtoonde lettergreep?

2.b) Difoonsynthese

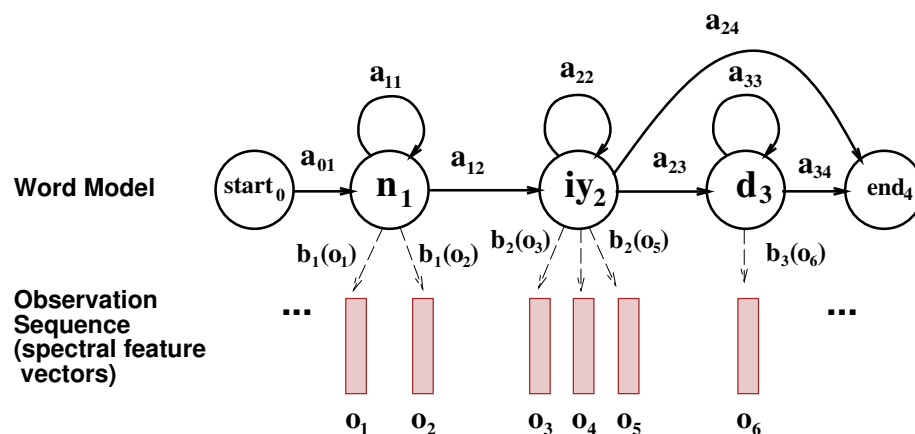
Bij veel tekst-naar-spraak synthesesystemen wordt gebruik gemaakt van difonen als bouwstenen voor het genereren van spraak.

- Wat zijn difonen en waarom denk je dat ze zulk een goede synthese kwaliteit geven?
- Wat zijn de belangrijkste beperkingen van difoonsynthese?

2.c) Difonen

- Welke difonen zijn nodig voor het genereren van de uiting: *Halt, taxi*?
- Welke difonen heb je nog extra nodig om ook het woord *halter* te kunnen genereren?
- Hoe zou je de bovenstaande difonen met de hand kunnen aanmaken uit een lijst met voorgelezen woorden of zinnen?

3 Hidden Markov Model ASR



Figuur 2: Versimpeld Hidden Markov Model van het Engelse woord *need*

3.a) Hidden Markov Model

Leg uit wat met de symbolen a_{ij} , $b_i(o_k)$ en o_k in de bovenstaande figuur bedoeld wordt

3.b) Berekening van parameters

Ga ervan uit dat je alle onderliggende state overgangen, S_{ij} , voor elk paar fonemen in een groot spraakcorpus kent.

Hoe kun je dan eenvoudig a_{ij} en $b_i(o_k)$ berekenen?

3.c) HMM training

Hoe train je a_{ij} en $b_i(o_k)$ als je alleen de fonemen van de opgenomen spraak kent?

Ga niet te veel in op de details.

3.d) Foneemmodellen

In bovenstaande figuur wordt elk foneem gemodelleerd met één enkele, simpele onderliggende HMM state. In de praktijk worden fonemen anders gemodelleerd, hoe?

Illustreer met een simpele figuur.

4 Dialoogsystemen

1. USR: We need to get the woman in Penfield to Strong
2. SYS: OK
3. USR: What vehicles are available?
4. SYS: There are ambulances in Pittsford and Webster
5. USR: OK. Use one from Pittsford
6. SYS: Do you know that Route 96 is blocked due to construction?
7. USR: Oh
8. USR: Let's use the interstate instead
9. SYS: OK. I'll dispatch the crew

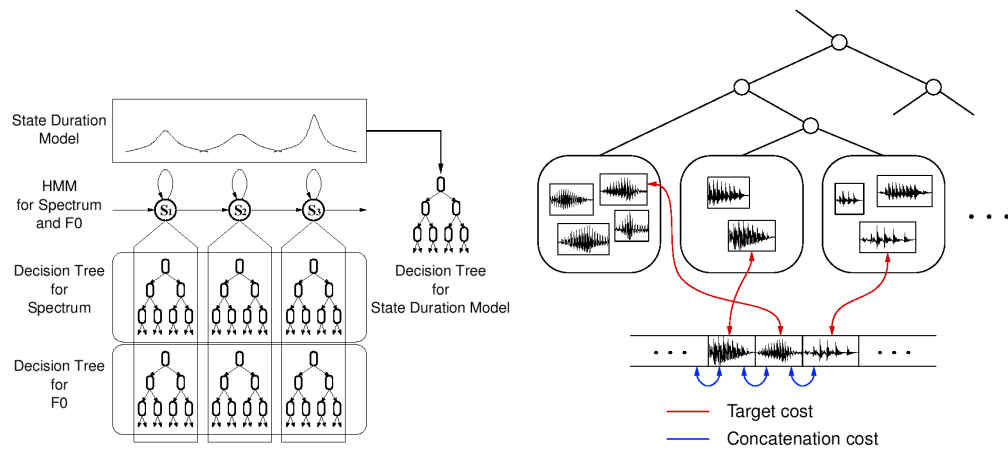
Figuur 3: Voorbeeld van een *disaster recovery* mens-machine dialoog.

USR: Gebruiker, SYS: Systeem (machine), Strong: *Strong Memorial Hospital*

Hierboven zie je een fragment uit een mens-machine dialoog. Dit fragment illustreert enkele aspecten van *automatic dialog management*.

- 4.a) Gebruiken de sprekers *acknowledgements*? Zo ja, waar?
- 4.b) Het systeem herkent de *intenties* van de gebruiker. Waar is dit evident?
- 4.c) Geef bewijzen dat het systeem *goal driven* is.
- 4.d) Wat is *barge-in*, en hoe functioneert dit in dit fragment?

5 HMM Text-To-Speech



Figuur 4: *Decision trees* voor context clustering en clustering-based unit selectie schema

Hierboven zie je een schematische weergave van het gebruik van clustering in Hidden Markov spraaksynthese. Dit is een vorm van *unit selection* synthese.

- 5.a) Units worden vooraf geclusterd op basis van context. Welke contextfactoren worden voor het clusteren gebruikt? Wat is het voordeel van clusteren?
- 5.b) Spectrum, F_0 , en duur worden apart gemodelleerd in HMM synthese. In HMM spraakherkenning wordt enkel het spectrum gemodelleerd. Wat is de reden van dit verschil?
- 5.c) Met HMM spraaksynthese kan een bestaande “stem” op eenvoudige wijze omgezet worden naar een nieuwe taal of een stem met andere sprekercharacteristieken. Leg uit hoe.
- 5.d) Wat is naast de flexibiliteit van (c), een andere belangrijk voordeel van HMM synthese boven klassieke unit-selectie?