

## 2<sup>e</sup> Deeltentamen Spraakherkenning en -synthese

Rob van Son

9-12 uur, 17 december 2000  
REC-P 2.27

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 deeltentamen 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. Reken op 5 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. Be-steed 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) Automatische toonherkenning

In opgave 7 hebben jullie een toonherkenner gemaakt. Het principe was dat je verschillende tonen genereert en via Dynamic Time Warping (DTW) de best passende zoekt. Hetzelfde als in het programma SpeakGoodChinese.

### 1.a) Dynamic Time Warping

Beschrijf in het kort hoe DTW werkt en hoe je deze techniek in bovenstaande opdracht kunt gebruiken. Ga uit van een keuze tussen twee tonen (stijgend en vlak).

### 1.b) Normalisering

Waarom moeten de toonhoogte en toonexcursie genormaliseerd worden om effectieve herkenning te krijgen?  
Hoe normaliseer je het best?

### 1.c) DTW over langere fragmenten

Is de DTW-methode die je hier en die in SpeakGoodChinese wordt gebruikt, uitbreidbaar voor uitingen die uit meer dan twee toonfragmenten achter elkaar bestaan?  
Beargumenteer.

### 1.d) Klankherkenning

In opgave 6 hebben jullie DTW gebruikt om cijfers te "herkennen". Hierbij werd als representatie de MFCC gebruikt. Deze MFCC's modelleren het spectrum en zijn gebaseerd op de uitvoer van een filterbank met 24 bandfilters. Krijg je betere resultaten als je het spectrum nauwkeuriger modelleert met bijvoorbeeld 48 in plaats van 24 bandfilters?

## 2) 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%.

$$WER = \frac{S + I + D}{W} * 100\% \quad (1)$$

Als de woordsequentie van de ingesproken letterreeks was: PQLKJIHGF, wat is dan, uitgaande van een optimale match,

- de WER bij een herkende output: PQMKJIIGFN, en
- wat bij een herkende output: LKJJIH?

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

---

### 3) Dialoogsystemen

#### 3.a) Stroomarchitectuur

Geef een eenvoudig schematisch overzicht van de architectuur van een *conversational agent*, ofwel een automatisch dialoogstroom. Een blokschema is voldoende.

#### 3.b) Hypothesen over conversaties

*The Practical Dialogue Hypothesis:*  
The conversational competence required for practical dialogues, while still complex, is significantly simpler to achieve than general human conversational competence.

*The Domain-Independence Hypothesis:*  
Within the genre of practical dialogue, the bulk of the complexity in the language interpretation and dialogue management is independent of the task being performed.

Als de bovenstaande twee hypothesen uit het artikel van Allen et al. waar zijn, wordt de ontwikkeling van automatische dialoogstroom sterk vereenvoudigd. Geef in het kort aan hoe deze twee hypothesen gebruikt kunnen worden bij de studie en het ontwerp van automatische dialoogstroom.

Gebruik als voorbeeld een *emergency rescue system* zoals besproken in het college en het artikel van Allen et al.

#### 3.c) Dialoogtechnieken

*Dialoogtechnieken*

- Finite-state Script
- Frame-based
- Set of Contexts
- Plan-based models
- Agent-based models

*Dialoog voorbeeld*

- 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

Links staat een lijst van technieken waarmee dialoogstroom gemaakt kunnen worden. Rechts een voorbeeld van een *emergency rescue dialog*. Geef voor iedere techniek in het linker lijstje aan of die bruikbaar is voor het voorbeeld.

Zo niet, waarom niet? (geef voorbeeld, houdt het kort)

Zo ja, waarom wel? (houdt het kort)

#### 3.d) Conversational Maxims of Grice

Geef de Maxims van Grice. Zoek voor elk van de maxims een voorbeeld uit dialoog hierboven. Als dat er niet is, maak dan zelf een voorbeeld.

---

## 4) Template based speech recognition

### *NORISC : Next generation template based Recognition for Interactive man-machine Speech Communication*

It is studied in this research program how a state-of-the-art approach to ASR, can be adapted, or even be replaced by an approach in which some of the implementation choices are each made better suited to describe the long-span speech processes of pronunciation variation. This research program focuses on three different domains: (1) the choice of acoustic features, (2) the choice of the recognition units and the attending challenge for the dynamic programming search, and (3) the choice of how to define new distance measures. In this manner, the research program aims at developing a new approach to acoustic decoding in ASR, advancing the state-of-the-art in HMMs, while keeping its advantages.

### 4.a) Uitspraakvariatie

Beschrijf hoe "klassieke" spraakherkenningssystemen, zoals OVIS, de problemen met variatie in de uitspraak van woorden beschrijven en oplossen. Gebruik als voorbeeld variatie in het woord *goedendag* ([xudəndɑx]).

### 4.b) Spraak bestaat niet uit *kralen aan een touwtje*

NORISC (zie boven) wil uitspraakvariatie modeleren met *demisyllables* (halve syllaben). Dat zijn syllabe-fragmenten die eindigen of beginnen halverwege de klinker: [start] wordt [sta] + [art] met snijpunten midden in de [a].

Leg uit hoe deze aanpak van uitspraakvariatie geïmplementeerd kan worden met *Template based Recognition* zoals die in het college behandeld is. Ga ervan uit dat niet alle woordvarianten in hun geheel in het trainingsmateriaal aanwezig zijn.

### 4.c) HMM implementatie

In de oorspronkelijke aanvraag wilde NORISC gebruik maken van HMM herkeners. Het idee was om *demisyllable*- ipv *foneem*-modellen te gebruiken. Iedere *demisyllable* wordt dan gemodelleer als een netwerk van "gewone" HMM foneemmodellen. Uitspraakvariatie kan dan gemodelleerd worden als meerdere paden door de substates (demisyllable modellen).

Beschrijf deze aanpak en vergelijk haar met de standaard aanpak van triphone modellen in de HMM herkenning. Illustreer met een figuur.

### 4.d) Dynamische spectra

Een van de belangrijkste voordelen van *Template based Recognition* is het modeleren van een natuurlijk verloop van het spraakgeluid. NORISC wil verschillende manieren uitproberen om de spectrale dynamiek van fonemen en foneemovergangen beter te modelleren, de zgn Structure Based Method. Leg aan de hand van voorbeelden uit automatische spraakherkenning en -synthese waarom die dynamiek zo belangrijk is voor de herkenning en hoe dit verschilt van standaard HMM herkenning.