

## Tentamen spraakherkenning en -synthese (2)

23 december 2010

Zet op iedere pagina je naam, je studentnummer en het volgnummer per pagina. Gebruik voor elke opgave (1-5) een apart vel. Als je voor 15.00 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 als je alle practicumverslagen voor dit onderdeel hebt voltooid en ze allemaal 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. Besteed niet te veel tijd aan afzonderlijke opgaven. Als een opgave te veel tijd vergt, probeer dan eerst een andere opgave.

Veel succes!

David Weenink

# 1 Dynamic Time Warping (DTW)

1. Leg kort uit wat DTW is en geef een paar voorbeelden.
2. Leg kort uit hoe we de DTW-afstand tussen twee geluiden kunnen meten. Vermeldt ook de tussenstappen.
3. Gegeven twee signalen  $s_1$  en  $s_2$  met een respectievelijke duur van 2 s en 4.5 s. Tijdens de DTW bepaling gebruiken we een distance matrix van dimensie  $n_1 \times n_2$ . We zoeken het optimale pad en we eisen dat begin- en eindpositie matchen, dat wil zeggen het optimale pad begint bij matrixelement (1,1) en eindigt bij  $(n_1, n_2)$ . Waarom heeft de constraint dat de helling van het optimale pad moet liggen tussen 2 en 1/2 bij deze twee signalen geen zin?
4. Stel we willen DTW gebruiken voor het herkennen van de drie woorden in een verzameling  $S_i$ . We hebben de volgende verzamelingen,  $S_1 = \{\text{tapper, topper, tipper}\}$ ,  $S_2 = \{\text{tap, top, tip}\}$ , en  $S_3 = \{\text{taptoestel, toptoestel, tiptoestel}\}$ . Welke verzameling kunnen we het beste gebruiken? Rangschik de verzamelingen naar geschiktheid en  *motiveer*  je antwoord.

## 2 Automatische spraakherkenning via Bayes

Centraal in de automatische spraakherkenning staat

$$\hat{W} = \operatorname{argmax}_{W \in L} P(W|O),$$

waarin  $\hat{W}$  de meest waarschijnlijke zin is,  $W$  een kandidaat zin uit taal  $L$  en  $O$  het waargenomen geluid.

1. Wat bedoelen we met deze formule?
2. Deze formule is niet rechtstreeks implementeerbaar. Geef aan hoe we met behulp van Bayes deze formule verder kunnen uitwerken.
3. Geef aan wat de verschillende componenten in de uitgewerkte formule betekenen.
4. Welke term in de uitgewerkte formule is “overbodig” en waarom?
5. In de spraakherkenning onderscheiden we o.a. *feature extraction*, *acoustic modeling* en *decoding*. Geef een korte beschrijving van elk.

### 3 Hidden Markov Model ASR

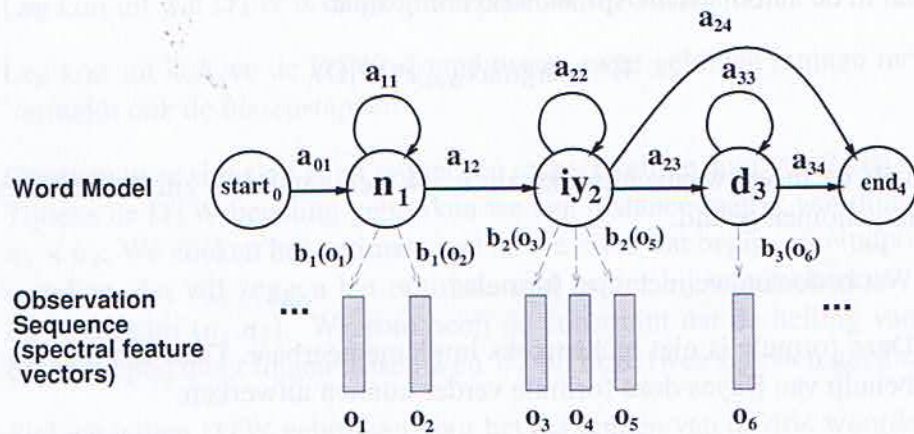


Figure 1: Versimpeld Hidden Markov Model van het Engelse woord *need*.

#### 3.1 Hidden Markov Model

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

#### 3.2 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?

#### 3.3 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.

#### 3.4 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 Formantsynthese

1. Waarom gebruiken we nog formantsynthese als unitsynthese al zo goed is?
2. Een formant kan als functie van de tijd geschreven worden als  $f(t) = e^{-\alpha t} \sin(2\pi Ft)$ , waarbij de demping  $\alpha$  en de formantfrequentie  $F$  positieve getallen zijn. Het *spectrum* van deze formantfunctie zal een piek vertonen ongeveer bij de frequentie  $F$ . De breedte van deze piek, de *bandbreedte*, zal afhangen van de demping  $\alpha$ . Hoe zullen de breedte én de hoogte van de piek veranderen als de demping verandert? Geef ook een kwalitatieve verklaring van de relatie tussen demping en bandbreedte. Illustreer eventueel met een tekening.
3. Welke vier onderdelen kunnen we onderscheiden in een akoestische formantsynthesizer zoals de Klatt(Grid) synthesizer.
4. De formantsynthese kan zowel parallel als serieel uitgevoerd worden. Welke parameters hebben we nodig om een formant te specificeren als we formantfilters in cascade gebruiken? Welke extra parameter(s) hebben we nodig voor parallel synthese?

exte  
= Amplitude?

## 5 Analyse van het spraaksignaal

1. Wat zijn mel frequency cepstral coefficients (MFCC)?
2. Beschrijf alle stappen om mel frequency cepstral coefficients te maken uit het (digitale) spraaksignaal. In welke van deze stappen treedt datareductie op?
3. Waarom gebruikt men de *mel frequency scale* en geen *linear frequency scale* in de spraakherkenning?
4. Waarom zou men in de spraakherkenning behalve de MFCC's van een frame ook nog de verandering van de MFCC's tussen frames gebruiken (de zogenaamde delta's en de delta delta's)?
5. Waarom is een afstandsmaat tussen twee MFCC representaties betrouwbaarder dan tussen twee gewone spectra (hint: denk aan fundamentele frequentie)?