Spraakherkenning

en

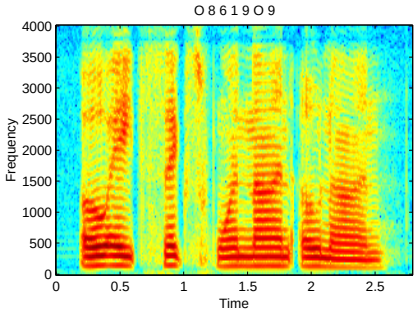
spraak

Onderwerpen

- spraaksignaalbewerking
- principe automatische spraakherkenningssystemen (ASR)

- omgaan met de complexiteit van de “echte” wereld
- de problemen van spraakherkenning
- cognitie en techniek
- conclusies

Spectrum van “O 8 6 1 9 O 9”



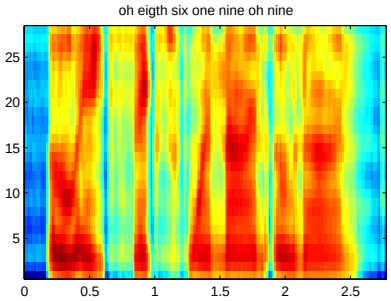
Spraak

- frequentie bijdragen
- veranderingen in de tijd
- vaak periodiek: (horizontale structuren)
- aperiodieke bijdragen (vaak vertikaal)

Klinkers

- harmonischen (horzontale structuren)
- formanten: complexen van harmonischen
- klinkeridentiteit: positie van formanten

Spraaksignaalbewerking



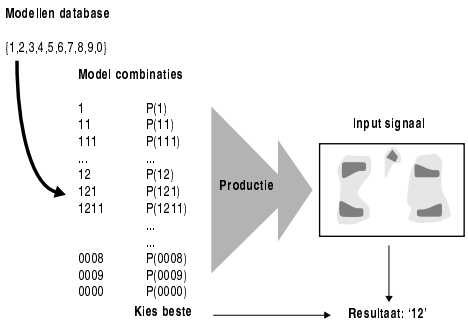
Woordidentiteit in formantpatroon:

- spectrale omhullende (MFCC)

Bulten in spectrale omhullende (vertikaal) corresponderen met formanten

Automatische spraakherkenning op basis van spectrale omhullende,

Computerspraakherkenning (1)



Een spraakherkenner is een systeem dat elke input afbeeld op de reeks van opgeslagen woordmodellen die het meest op de input lijkt.

In feite wordt van elk van elke combinatie van woordmodellen berekent wat de kans is dat de combinatie de input produceert: het beste model is het herkenningsresultaat.

Computerspraakherkenning (2)

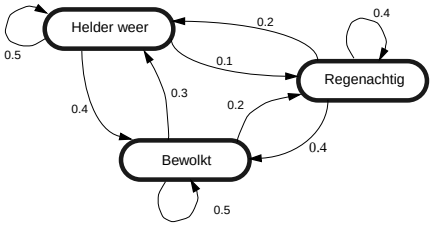
Uitgangspunten:

- de input wordt gerepresenteerd door een combinatie van *statistische* modellen
- de meest waarschijnlijke woordreeks (gegeven de akoestische evidentie) is voldoende vaak de correcte woordreeks

Basisgereedschap

- statistiek, signaalanalyse (geen cognitie wetetenschap)
- akoestische woordmodellen $P(y|w)$: Hidden Markov Modellen (HMMs)
- statistische taalmodellen $P(w)$: de kans dat woorden in de context van voorgaande woorden voorkomen

Basis formule: $\hat{w} = \arg\max_w \left\{ \frac{P(y|w)P(w)}{P(y)} \right\}$



Markov Modellen

Voor modellering (tijd)reeksen

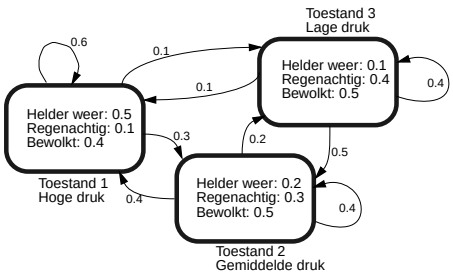
- toestanden
- overgangswaarschijnlijkheden

Volgende toestand alleen afhankelijk van huidige toestand

Meest waarschijnlijke reeks:

... - bewolkt - bewolkt - bewolkt - ...

Problemen bij veel of variërende toestanden
Oplossingen: *Hidden Markov Models* (HMM)



Hidden Markov Modellen

- nu ook waarnemingswaarschijnlijkheden

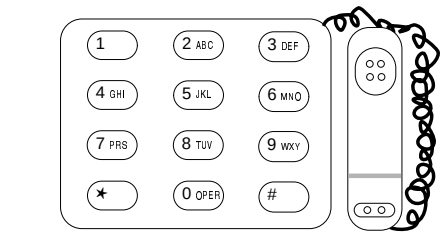
In elke toestand kan elke waarneming gedaan worden, maar steeds wel met een andere kans

- dus toestandsreeks onbekend

Meest waarschijnlijke reeks:

... - bewolkt - bewolkt - bewolkt - ...

Voor elk klimaat (klank) een ander HMM



Toetsen	25929715665128184312744481744481631543O
Mogelijke letters	ajwawp jmmj at tgd apgggt pgggt md jgdd.
	bkxbxr knnk bu uhe brhhhu rhhhu ne khee.
	clycys lool cv vif csiiiv siiiv of liff.
Boodschap	always look at the bright sight of life

Telefoonanalogie

- Start: meest waarschijnlijke keuzes aj, ak, al en alw, alx, aly
- *Backtracking* na "jon" i.p.v. 'loo"
- Zekerheid aan einde zin
- Zekerheid na testen alle mogelijkheden
- Gebruik betekenis: "*right*" vs "*sight*"

Uitbereiding telefoonanalogie

- spraakherkenning volledig afhankelijk van akoestische informatie
- i.v.m. met coarticulatie 40 tot 10000 klankmogelijkheden (in plaats van 10)
- er is geen duidelijke segmentatie, elke klank kan elk moment beginnen

Beam search (*beam* = straal van zoeklicht)

- doorzoek alleen de meest waarschijnlijke mogelijkheden

Het horten en stoten gedrag van dicteerprogramma's

- Na ....
- Naar raam...
- Naar raam leiding...
- Naar aanleiding van nu ver zoek...
- Naar aanleiding van uw verzoek om...

Keuzes

Elke keuze voor een beter systeem geeft grotere problemen

- Sprekerafhankelijk of spreker onafhankelijkheid
- Losse woorden, verbonden of continue spraak
- Keuze van de woorden in het systeem
- Domeinkennis
  - domein afhankelijke grammatica's
- Dialoog management

- Omgaan met onverwachte input
  - naieve gebruikers
  - (achtergrond) ruis, slechte verbinding
  - meerdere sprekers, muziek
  - accenten, kinderen
  - nieuwe woorden

----====+++++====----

Real-world complexity

Eigenschappen van onze leefomgeving:

- onvoorspelbaar
- oncontroleerbaar
- onoverzienbaar
- onbegrensde complexiteit

Echter ook:

- in hoge mate voorspelbaar
- grotendeels invariant
- onderhevig aan (fysische) wetten

Eigenschappen perceptief systeem:

- opgeslagen kennis
  - maakt voorspelling mogelijk
  - representeert "invarianties"
  - (fysische) regelmatigheden
- onvoorspelbare input vereist constante vergelijking van input met opgeslagen kennis (een *mismatch* trekt aandacht!)

Vraagje

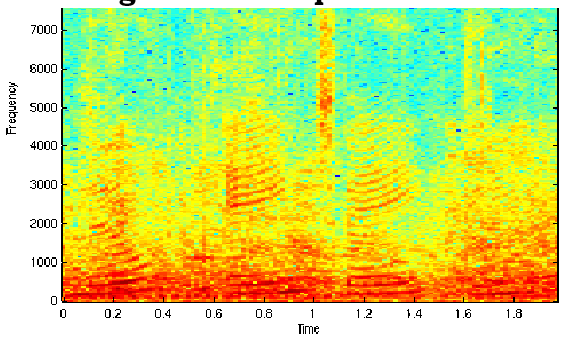
Er is precies 1 type geluidsbron dat (op basis van geluid alleen) nooit correct te herkennen is. Welke?

Een ideale geluidsinstalatie kan geluid zo goed reproduceren dat niemand meer het verschil kan horen tussen de het origineel en de reconstructie: een ideale geluidsinstalatie is daardoor onherkenbaar!

Een ideale geluidsinstalatie heeft geen hoorbare *fysische* beperkingen.
Alle andere geluidsbronnen hebben die wel.

Conclusie: we herkennen, classificeren/en scheiden geluidsbronnen op basis van een vergelijking met kennis over fysische beperkingen van geluidsbronnen!

De signaal in ruis paradox



Onbekend signaal

- aantal bronnen?
- soort signalen?
- start en einde van bijdragen?
- wat hoort bijelkaar?
- wat is relevant?

Paradox

- correcte herkenning na correcte selectie
- correcte selectie na correcte herkenning

Robuustheid - Betrouwbaarheid

Gekoppelde vragen:

- hoe kan het menselijke spraakproces in zoveel situaties adequaat functioneren?
- hoe kan een automatische spraakherkenner *real-life input* betrouwbaar verwerken?

Antwoorden

1) Oplossing signaal in ruis paradox:

- integreer selectie en herkenning

  Herkenning op basis van de *beste* en een *voldoende* goede overeenkomst.

2) Ga uit van optimaal zwakke (algemeen geldende) basis aannames

- signaalcomponenten bestaan uit: *onset - continue ontwikkeling - offset*

Basisaannames automatische spraakherkenning

Basis aannames zijn zeer sterk (d.w.z. in het algemeen niet valide):

- de input is spraak van de juiste taal en kan correct herkent worden
  - *eerst selectie daarna pas herkenning*
  - muziek, achtergrond geluid
  - andere talen, dialecten
- de input kan correct gerepresenteerd worden door een combinatie van getrainde *statistische* modellen
  - te weinig trainingsdata -> geen goede herkenning
  - onbekende woorden, spontane en/of spraak, “emotionele” spraak
- de meest waarschijnlijke woordreeks (gegeven de akoestische evidentie en het taalmodel) is de correcte woordreeks
  - domeinspecificiteit

Natuurlijke spraakverwerking

Evolutionaire druk leidt tot een optimalisatie van belangrijke biologische processen

Postulaat1

- het auditief systeem functioneert adequaat in zoveel mogelijk akoestische omgevingen (d.w.z. zo vaak mogelijk)

Postulaat 2

- spraak is een communicatievorm die in zoveel mogelijk akoestische omstandigheden leidt tot een correcte informatie uitwisseling

De meest *informatieve* linguistische kenmerken (b.v. formanten) zijn, *door gebruik te maken van continuiteit in het signaal*, zeer robuust vast te stellen.

- auditief systeem bewaart continuiteit!

Geluidsvoorbeelden

Voorbeelden auditieve integratie en segregatie, o.a. op basis van continuiteit (Bregman, *Auditory Scene Analysis*):

- Segregatie door toonhoogte verschil
  - 3: frequentieverschil groot vs klein
  - 5: "*Mary had a little lamb*"
  - 6: Teleman fluitconcert
- Groepvorming door continuiteit
  - 12: verbonden vs los
- Groepvorming door "*common fate*"
  - 19: temporele ontwikkeling
  - 24: zang: toon, complex, vibrato
- Spraakachtige evolutie
  - 23: sine-wave speech: "*Please say what this word is: sill, shook, rust, wed, pass, lark, jaw, coop, beak*"

Conclusies (1)

Automatische spraakherkenning is in triviale situaties mogelijk:

- 1 spreker
- vast applicatiedomein
- geen (complexe) achtergrondgeluiden

Toepassingssituaties zijn zelden triviaal, want mensen leven in complexe omgevingen

Centraal kenmerk menselijke perceptie:

- (schijnbaar moeiteloos) functioneren in complexe, variabele situaties

Taak kunstmatige intelligentie/ cognitiewetenschap:

- ontwerp systemen die om kunnen gaan met zeer complexe, variabele en onbekende input

Conclusies (2)

Uitgangspunt robuustheid

- maximale robuuste herkenning op basis van maximaal zwakke basisaannames
- ontwikkel/bouw eerst algemene aanpak, en maak deze taakspecifiek (dus niet omgekeerd, zoals bij ASR)

Signaal in ruis paradox:

- herkenning en selectie in één proces

Perceptie:

- constante vergelijking tussen input en verwachtingen op basis van geleerde kennis: b.v. (fysische) regelmatigheden

Herkenningsresultaat:

- de beste herkenningshypothese die *voldoende* goed past

Zinvolle, ruisrobuuste spraakherkenningsprodukten zijn pas mogelijk na integratie van cognitiewetenschappelijke kennis in herkenningsproducten