



Spraakherkenning voor transcriptie van middeleeuwse teksten

Patrick Wambacq

K.U.Leuven/ESAT

PSI-Speech



Transcriptie d.m.v. spraakherkenning

- werkwijze:
 - een geoefend persoon met een goede stem leest de teksten voor; consistentie van uitspraak gewenst
 - het audiosignaal wordt offline door een spraakherkenningsysteem gestuurd
 - de output van de spraakherkenner is niet perfect en moet/kan manueel geverifieerd worden
 - als indexering volstaat i.p.v. transcriptie dan is de kwaliteit van de output minder kritisch



Spraakherkenning

- systeem voor herkenning van continue spraak met groot vocabularium wordt getraind voor de doeltaal en rekening houdend met de akoestische omgeving
- in gunstige omstandigheden wordt een nauwkeurigheid van 95% of hoger bereikt (voor hedendaagse talen en voor een dicteertaak)



Spraakherkenningsproblemen

- dikwijls minder goede omstandigheden: akoestische omgeving (telefoon, ruis, nagalm, ...), overlappende spraak, sprekervariabiliteit (allochtonen, accenten/dialecten, kinderen, ouderen, slechte articulatie, ...), spontane spraak (disfluencies), ... → niet echt van toepassing hier
- nieuwe namen duiken voortdurend op: lexicon en taalmodel moeten uitgebreid worden → hier wel van toepassing
- punctuatie in de output ontbreekt



Problemen met Middelnederlands

- ongekende woorden: andere woordenschat dan huidige spraakherkenner
- geen aangepast taalmodel (N-gram: waarschijnlijkheid van huidig woord gegeven N-1 voorgaande woorden)
- verschillende uitspraken mogelijk
- geen standaard orthografie
- deze problemen stellen zich des te meer voor namen
- hoeveel verschillende soorten Nederlands?



Mogelijke oplossingen

- verzameling van (heel) veel gelijkaardige teksten zodat een taalmodel kan opgesteld worden; probleem: tekstnormalisatie
www.dbnl.org ?
- fonetisch lexicon nodig: gebruik een aangepaste grafeem naar foneem omzetter; uitspraak van de woorden moet vastgelegd worden
- normalisatie van het eindresultaat: wegfilteren van disfluencies



Demos

- herkenning van voorgelezen middeleeuwse tekst
- transcriptie van nieuwsuitzending op radio/tv



Vragen?