Monitoring van een ASR-systeem

Het gebruik van interne confidencewaarden en teletekstondertiteling om de Word-Error-Rate van een herkenning te voorspellen.

- Universiteit Twente -

- Faculteit Elektrotechniek, Wiskunde en Informatica -

Auteur Laurens Brouwer

Datum Augustus 2006

Colofon

Universiteit Twente

Faculteit EWI

Vakgroep HMI

Drienerlolaan 5

7522 NB Enschede

www.ewi.utwente.nl

Eerste begeleider: Ir. M.(Marijn) A.H.Huijbregts

Tweede begeleider: Dr. R.(Roeland) J.F. Ordelman

Derde begeleider: Prof. dr. F.(Franciska) M.G. de Jong

Vierde begeleider: Dr. A.(Arjan) J. van Hessen

Auteur

L.J.Brouwer

Campuslaan 55-102

7522 NK Enschede

l.j.brouwer@student.utwente.nl

Samenvatting

Monitoring in de context van automatische spraakherkenning(ASR) houdt in dat de prestatie van de herkenner bijgehouden wordt zonder tussenkomst van een mens. De monitoring mét tussenkomst van een mens gebeurt door middel van handmatige transcriptie, waarin de persoon aangeeft wat gezegd is. Het meten gebeurt door herkenning en transcriptie te vergelijken. De Word-Error-Rate die zo berekend wordt is de maat voor de prestatie van de herkenning. Bij monitoring zónder tussenkomst van de mens worden interne likelihoods en teletekstondertiteling gebruikt om een maat voor de prestatie te bepalen.

Het systeem of kader waarbinnen dit onderzoek gedaan wordt is Spoken-Document-Retrieval. Op de Universiteit Twente is een website waar de dagelijkse journaaluitzendingen van 20.00 uur herkend worden. De gebruiker kan in de herkenning zoeken en vervolgens de videofragmenten bekijken. Monitoring moet aangeven hoe goed een zin, of heel journaal herkend is. Als blijkt dat de kwaliteit van de herkenning van enkele zinnen, of van een heel journaal te laag is, kan alarm geslagen worden.

Voor hertraining of adaptatie is akoestische data samen met tekst (herkenning) nodig. Monitoring kan helpen om automatisch de beste data te selecteren om hierbij de beste resultaten te bereiken.

Voor monitoring zijn twee methoden onderzocht, één wordt “interne confidence measure” genoemd, de ander “externe confidence measure”. De interne versie is gebaseerd op de likelihoods die de herkenner geeft van elk herkend foneem, de externe versie gebruikt de teletekstondertiteling als een substituut voor de handmatige transcriptie.

De interne confidence-measure is al vaker onderzocht. Er worden in dit verslag twee methoden besproken die voor een systeem als dit, met continue spraak en voor nieuwsuitzendingen, bruikbaar zijn. Deze twee methoden en de vele varianten hierop zullen verder worden onderzocht door experimenten te doen, waarna de beste methode voor interne confidencewaarden bepaald kan worden.

De teletekstondertiteling is soms een letterlijke weergave van wat er in het journaal te horen was, maar meestal zijn er kleine verschillen. Er zijn zelfs gevallen waarin de ondertiteling volledig mist. Daarnaast is de koppeling tussen ondertiteling en herkenning een probleem. De ondertiteling loopt niet perfect mee met de spraak. Er is onderzocht in hoeverre, ondanks deze problemen, gebruikt gemaakt kan worden van teletekstondertiteling om te monitoren.

Het laatste onderwerp is het toepassen van monitoring voor training. Er is een experimentele toepassing gemaakt die, op basis van de confidence-waarden, een selectie maakt van de beste data. Dit onderzoek moet inzicht geven in de bruikbaarheid van de confidencewaarden voor toepassingen.

De experimenten hebben aangetoond dat de monitoring gemiddeld gezien goed aangeeft hoe goed de herkenning is. Alle methoden geven aan dat er een duidelijke lineaire correlatie bestaat tussen de bepaalde scores en de daadwerkelijke Word-Error-Rate. Bij het training-experiment blijkt dat selectie van de beste zinnen, op basis van de confidence-measures, goed werkt.

Abstract

Monitoring in automatic speech recognition (ASR) implies that the Word-Error-Rate of the recognition can be predicted without human intervention. Human intervention in this case means that a human will listen to the speech and write down what is said. This is called transcribing. Measuring is then done by comparing the recogniser-output with the transcription. The Word-Error-Rate(WER) is **the** measure of how well a recognition was done. The proposed monitoring will use internal likelihoods and closed caption to calculate new measures that estimate the WER.

This monitoring is developed for the Spoken-Document-Retrieval system that is developed at the University of Twente. Here a web-portal is online where the daily 20.00 'o clock broadcast-news is automatically recognised. The user can then search in the textual representation of the audio, and view the video of the corresponding fragment. For this system, monitoring should indicate the quality of parts, or all of the recognition. If monitoring shows that some parts are badly recognised, actions can be taken manually to fix the problem.

The two methods are called "internal confidence measure" and "external confidence measure". The internal measure wil use the acoustic phoneme-likelihoods that are given in the recognizer-output to calculate a value that can estimate the WER. Two methods that were found in publications were investigated. Some modifications were researched for use in this system, after which experiments were done to determine the most useful method.

The external measure uses closed caption as a substitute for the manual transcription. Closed caption is sometimes literally equal to what is spoken. In most cases however there are some differences or huge differences. Research is done to determine an approach that is usable in general, despite the difference in quality between the individual captions. By using the closed caption as if it were a manual transcription, it should be possible to calculate a WER in the normal way, and then estimate the real WER afterwards.

One possible application of monitoring is selecting the best parts of the recognition for training or adaptation of the acoustic model. A few experiments have been done to reflect on the usefullness of monitoring in this form.

The experiments have showed that, in general, monitoring can estimate the WER quite well. Due to the differences in quality of internal- and external confidence mearures, the WER of individual recognition-segments cannot be determined very reliably. Both confidence measures indicate a clear linear correlation between the measure and the WER. The training-experiments have showed that when automatically selecting the segments with the lowest estimated WER, according to monitoring, in general really the best segments are selected.

Voorwoord

Geachte lezer,

dit verslag is het resultaat van mijn afstuderen bij de vakgroep Human Media Interaction aan de Universiteit Twente. Ik ben sinds september 2005 tot nu, eind augustus 2006 bezig geweest met de opdracht “On-line monitoring”, al gauw omgedoopt naar “monitoring van een ASR-systeem”, wat uiteindelijk ook de titel is van dit verslag.

Ik ben op de spraak&taal-afdeling terechtgekomen bij mijn stage, eind 2004. In 2005 was er nog sprake van een afstudeeropdracht in Engeland, maar uiteindelijk is het Enschede geworden.

De opdracht die ik heb gedaan betreft het automatisch analyseren van de output van de spraakherkenner. Ik heb de herkenning van enkele 20.00 uur journaaluitzendingen bestudeerd en laten analyseren. De ontwikkelde methoden bepalen automatisch hoe goed de herkenning is geweest, met behulp van interne gegevens van de herkenner, en externe gegevens in de vorm van de teletekst-ondertiteling (via pagina 888). Ik heb helaas geen implementatie meer kunnen doen. Het is altijd leuk om te zien dat een methode in de praktijk werkt, maar na een jaar onderzoeken moet er een keer een punt achter gezet worden. Desalniettemin ben ik blij met de resultaten. Het lijkt er op dat een implementatie (door iemand anders) goede resultaten op zal leveren.

Ik heb bij mijn onderzoek veel hulp gekregen van mijn begeleider Marijn Huijbregts. Bij deze wil ik hem daarvoor bedanken. Daarnaast bedank ik Roeland Ordelman, Franciska de Jong en Arjan van Hessen voor het deelnemen aan mijn afstudeercommissie en de vele tips en correcties tijdens het onderzoek en het schrijven van deze scriptie.

Daarnaast bedank ik Marieke Beld en Vincent Weisscher voor de hulp bij het schrijven van deze scriptie.

Ik wens u veel plezier bij het lezen van deze scriptie.

Laurens J. Brouwer

Enschede, augustus 2006

Inhoudsopgave

1. Inleiding

In deze afstudeerscriptie worden twee methoden gepresenteerd die zonder menselijke tussenkomst herkenningen van een automatische spraakherkenningsysteem (ASR) evalueren. Dit evalueren wordt monitoring genoemd. In dit hoofdstuk wordt het systeem beschreven waarbinnen de opdracht is ontstaan, en wat het doel van de onderzoeken is.

1.1. *De opdracht*

Een van de onderzoeksonderwerpen van de afdeling spraak en taal van de Universiteit Twente (UT) is spoken document retrieval (SDR). Bij SDR wordt de spraak in geluidsbestanden herkend en de tekstuele representatie hiervan (de herkenningsresultaten) in een database geplaatst, waardoor gezocht kan worden in de gesproken tekst. De geluidsbestanden kunnen zowel zelfstandige bestanden zijn (radio-opnamen, interviews, telefoongesprekken) als onderdeel zijn van Multi-media documenten (TV, film, webcasts etc.). SDR faciliteert de indexatie van, en het zoeken naar audio- en videofragmenten. (Huijbregts, Ordelman et al. 2005)

Het SDR-systeem is ontwikkeld voor de web-portal van het Willem Frederik Hermans Instituut (WFH), dat op haar website de mogelijkheid biedt om te zoeken in boeklezingen, interviews en reportages. De gebruiker typt een trefwoord in, waarna van de zoekresultaten het bijbehorende audio- of videofragment afgespeeld kan worden. Het bestaande systeem zal worden uitgebreid met steeds meer multimedia-data. Het huidige systeem werkt betrekkelijk goed omdat de kwaliteit van de data goed is. Het zou kunnen dat in de toekomst toegevoegde data van mindere kwaliteit is dan de huidige, waardoor de herkenning slechter zal zijn, en het systeem minder goed zal werken. Om de kwaliteit van de herkenning op de lange termijn te volgen, zonder dat daarbij veel handmatig werk bij komt kijken, is automatische monitoring nodig. De oorspronkelijke afstudeeropdracht (zie bijlage C) had als doel de mogelijkheden tot monitoring te onderzoeken, en implementatie van monitoring binnen het WFH-systeem te realiseren.

Op de UT is nog een SDR-systeem online (het Broadcast-News systeem ofwel BN), dat in de volgende paragraaf besproken wordt. De afdeling HMI doet veel onderzoek met de BN-data, omdat het duidelijke spraak is, er mogelijkheden tot updating zijn en goede beschikbaarheid van veel gerelateerde data is. Daarom heeft het aanvankelijke onderzoek voor deze scriptie zich hierop gericht. Naar aanleiding van de bevindingen van dit onderzoek (zie komende twee paragrafen), zijn naast de hierboven genoemde lange-termijn-monitoring nieuwe toepassingen van monitoring bedacht. Er is besloten het verdere onderzoek te richten op deze nieuwe toepassingen, en de experimenten niet op data van het WFH-systeem, maar die van het BN-systeem te doen. Door het onderzoek op BN-data te doen is het ook mogelijk geweest de ondertiteling-onderzoeken te doen.

1.2. *Broadcast news demo*

Het SDR-systeem dat bij de vakgroep HMI online is, geeft de gebruiker de mogelijkheid om in nieuwsberichten te zoeken van het 20.00 uur journaal. Dit is het “broadcast-news-demo”-systeem[1]. Naast het zoeken in de herkende textuele representatie (de herkenning) van het journaal, biedt het daarnaast de mogelijkheid te zoeken in de teletekstondertiteling van hetzelfde journaal.

[1] Te vinden op http://hmi.ewi.utwente.nl/showcase/broadcast-news-demo (3 juli 2006)

- Er kan (klassiek) worden gezocht in het journaal van de laatste 7 dagen, ofwel op basis van de teletekst ondertiteling, ofwel op basis van de spraak die is omgezet naar tekst met behulp van een Nederlandse spraakherkenner. De gehele verwerking geschiedt automatisch: het videobestand wordt automatisch geknipt in segmenten die gebaseerd zijn op onderwerpwisseling (bij teletekst) of sprekerwisseling (bij spraakherkenning).

Tabel 1.1 De beschrijving van de zoekfunctie volgens de journaal-demo van HMI. De gebruiker kan de gevonden videofragmenten aanklikken uit een lijst met resultaten, en daarna bekijken.

In figuur 1.2 is een van de zoekresultaten te zien. Hierin wordt een gedeelte van de herkenning van een videofragment getoond.

En na een beraad van **kabinet** kan ik zeggen dat ik unaniem de conclusie zijn gekomen dat er niet aanvaarde motie geen gevolgen heeft voor de functioneren van een **kabinet** dat de kam neergelegd in een brief aan de tweede Kamer en die zegt dat nader toelichten.. ... »

Tabel 1.2 Een van de resultaten van de zoekactie op de term "kabinet". Dit is een deel uit de herkenning van het journaal van donderdag 29 juni '06.

Door de automatische verwerking is er geen controle over de juistheid van de zoekactie. Het kan gebeuren dat het trefwoord (in dit geval "kabinet") wel in de herkenning voorkomt, maar niet echt is gezegd. Zoals in de figuur is te zien zijn er grammaticale fouten in de zinnen, wat er waarschijnlijk op duidt dat enkele woorden fout herkend zijn. De context geeft aan dat het woord "kabinet" waarschijnlijk goed herkend is, dus deze zoekactie is geslaagd.

1.3. *Performance measurement*

De audiobestanden worden automatisch in segmenten opgedeeld van korte stukjes, liefst één of enkele zinnen. De evaluatie van de herkenning gebeurt door de audio per segment door een mens te laten beluisteren en annoteren. De annotatie wordt ook wel transcriptie genoemd. Hierna wordt de (automatische) herkenning vergeleken met de (handmatige) transcriptie ter evaluatie. Het maken van transcripties kost veel tijd, en is op lange termijn niet te doen. Monitoring is echter wel gewenst, en zelfs noodzakelijk vanwege de volgende problemen.

- Op lange termijn kan de spraak en woordgebruik in het nieuws veranderen (denk aan andere nieuwslezer, nieuwe woorden). Om in de toekomst te kunnen beoordelen in hoeverre het systeem nog werkt is dus noodzakelijk dat dit gemonitord wordt. Dit is vergelijkbaar met het probleem dat in de opdracht gegeven is.
- Binnen huidige journaals gebeurt het regelmatig dat er segmenten (stukjes) zijn die telefoongesprekken, gesprekken in andere taal of muziek bevatten. Op dit moment worden deze "gewoon" herkend door de Nederlandstalige spraakherkenner. Deze herkenning is echter niet bruikbaar voor een systeem als dit, en zou niet in het systeem mogen blijven.

Deze problemen geven aan dat monitoring per segment moet gebeuren. Het WFH-systeem vereiste slechts monitoring per gehele audiobron, wat equivalent zou zijn met een hele journaaluitzending in deze toepassing.

1.4. *Toepassingen*

In deze paragraaf zijn enkele mogelijke toepassingen van monitoring in het kort beschreven. De basis van alle toepassingen is de analyse per segment, met als resultaat een schatting van de prestatie oftewel WER (Word-Error-Rate, zie 2.2.2).

Monitoring-tool

De onderzochte methoden “interne confidence” en ondertiteling-referentie, kunnen beiden een getal opleveren waaruit de prestatie van de herkenning kan worden afgeleid. Deze getallen kunnen bijvoorbeeld in grafieken worden gezet waardoor van het journaal een overzicht wordt gegeven van goede en minder goede herkenningen. Deze grafieken kunnen de beheerder helpen bij het gericht werken aan de problemen, zoals die in 1.2 genoemd zijn. De tool zélf, met grafieken, is niet gemaakt, maar de monitoringwaarden kunnen wel berekend worden per segment van het journaal.

Rangschikking

Naast de basisapplicatie “monitoring” is de analyse van segmenten eventueel bruikbaar voor de **rangschikking** van de zoekresultaten. Betere herkenningen hebben hogere scores en komen hoger in de rangorde te staan, zodat de gebruiker zekerder is dat zijn zoekterm voorkomt in het stuk journaal. Deze toepassing is niet verder onderzocht.

Training

Hertraining of adaptatie gebeurt door de audio samen met de beste delen van de herkenning opnieuw aan te bieden aan de herkenner, zodat deze een nieuw of aangepast akoestisch model kan maken. Hoe meer data, hoe meer de modellen kunnen worden aangepast, maar als de modellen worden getraind op slechte herkenningen, zal de herkenning ook slechter worden.

Zowel kwaliteit als hoeveelheid van die data heeft invloed op de verbetering van de prestaties. De onderzoeksvraag voor training is dan ook: “Welke samenstelling van hoeveelheid en kwaliteit kan de beste verbetering teweeg brengen, als uitgegaan wordt van imperfecte (input-) data?”. De monitoring wordt gebruikt om de beste data te selecteren. De gemonitorde selecties zullen waarschijnlijk lage WER hebben, maar omdat het niet gecontroleerd wordt, is dit niet zeker. De resultaten van dit experiment zeggen daarmee ook iets over de kwaliteit van de monitoring. Het onderzoek is niet zozeer gericht op de techniek achter de hertrainingsmethoden, maar op de rol van monitoring in het selecteren van de juiste data.

1.5. *Doel afstudeeronderzoek*

In de paragrafen 1.3 en 1.4 zijn de “nieuwe” problemen en toepassingen besproken die ertoe hebben bijgedragen dat dit onderzoek zich heeft verplaatst van het WFH-domein naar het BN-domein. Hieronder wordt het nieuwe doel van het onderzoek beschreven.

Dit afstudeeronderzoek richt zich in eerste instantie op het onderzoeken van de mogelijkheden van monitoring. Er moeten methoden ontwikkeld kunnen worden die binnen het broadcast-news-systeem toegepast kunnen worden. Met deze methoden kan dagelijks een monitoring van het herkende journaal bepaald worden. Daarnaast zal onderzocht worden in hoeverre deze monitoring kan helpen voor het verbeteren van het systeem op lange termijn, het *trainen*.

1.6. *Herkenner*

Voor de herkenningen van het journaal, in het systeem van het broadcast-news, wordt gebruik gemaakt van de herkenner *Sonic(Pellom 2001)*. Dit is een herkenner die in Colorado, VS, gemaakt is. Het is een goede herkenner, maar weinig toegankelijk. Voor het monitoren zijn veel interne gegevens nodig, meer dan aangeboden wordt. De herkenner die uiteindelijk gebruikt is voor het ontwikkelen van de monitoring is *Shout(Huijbregts 2005)*, die ontwikkeld is op de UT. De prestaties van deze herkenner zijn iets minder dan *Sonic*, maar bijna alle gewenste uitvoer kan beschikbaar worden gemaakt.

1.7. *Aanpak monitoring*

Een herkenningsbestand bestaat uit segmenten van verschillende grootte. De segmentgrenzen kunnen zinnen zijn, of een aantal zinnen die bij elkaar horen, volgens automatische of handmatige segmentatie. Het onderzoek naar monitoring richt zich niet direct op een applicatie, maar op het bepalen of voorspellen hoe goed afzonderlijke segmenten herkend zijn.

Voor het monitoren van het systeem zijn twee methoden onderzocht die in zekere zin iets zeggen over de herkenningen. Het zijn de analyse van interne confidencewaarden (gegevens die de herkenner oplevert), en het vergelijken met externe referenties (de teletekstondertiteling).

Interne confidence

Bij het herkennen van de audio wordt voor elke herkende klank een akoestische likelihood gegeven. Dit is een interne waarde die aangeeft wat de kans is dat die klank op dat moment de juist herkende klank is. Na automatische analyse van de akoestische likelihoods kan een interne confidence-measure voor de hele zin bepaald worden. Een zin met hoge interne confidence-measure (CM) wordt geacht goed herkend te zijn.

Teletekstondertiteling

Er is binnen HMI een nieuw soort confidence bedacht, gebruik van teletekstondertiteling als externe referentie. Van het 20.00 uur journaal worden de ondertitelingen al gebruikt om mee te zoeken in de broadcast-news-demo. Deze ondertitelingen kunnen dienst doen als referentie daar ze meestal veel lijken op wat er echt gezegd is. Als een herkende zin lijkt op de ondertiteling ervan, is de kans groot dat hij goed herkend is.

1.8. *Dit verslag*

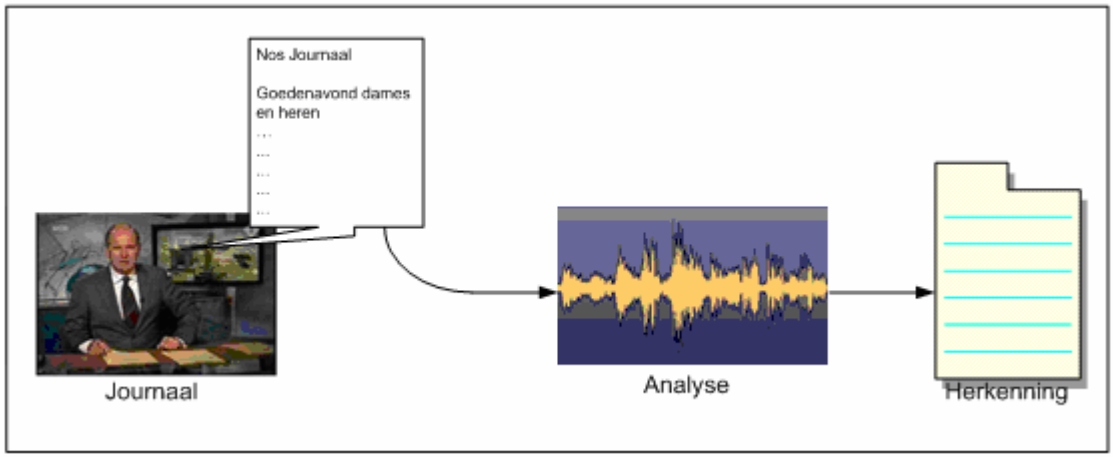
De hierboven aangegeven onderzoeken, toepassingen en experimenten zijn als volgt verdeeld over de hoofdstukken. Eerst is er een algemeen hoofdstuk over spraakherkenning. Het verslag is vrij theoretisch, vandaar dat voor de buitenstaander enkele begrippen en principes uitgelegd worden. De onderzoeken naar verschillende monitoringsmethoden worden besproken in de hoofdstukken 4 en 5, door literatuuronderzoek en vooral experimenten en statistische analyse. Daarna wordt de toepassing van van monitoring voor training besproken.

2. Spraakherkenning

In de volgende hoofdstukken wordt veel vakspecifieke kennis gebruikt die niet bij iedere lezer paraat is. Voor de lezer die niet op de hoogte is van de theorie achter de spraakherkenning wordt in dit hoofdstuk kort een overzicht gegeven van de werking van een spraakherkenner, en hoe de herkenning vervolgens verwerkt wordt.

2.1. *Theorie*

In tabel 2.1 wordt op het hoogste niveau de spraakherkenning uitgebeeld.



Tabel 2.1 schematische weergave van de spraakherkenning. De audio van het journaal wordt digitaal geanaliseerd, en uiteindelijk naar tekst omgezet.

De spraak moet gezien worden als audiosignaal. Dit signaal (de geluidsgolf) heeft een aantal fysieke kenmerken waardoor het meetbaar is voor de computer, namelijk de lengte (of eigenlijk de frequentie) en de amplitude (van de golf). Een spraakherkenner zet deze fysieke kenmerken (de feature vectors) om naar tekst. Dit is globaal hoe een spraakherkenner werkt.

In dit hoofdstuk wordt de weg van audio naar herkenning besproken. Eerst wordt uitgelegd hoe een herkenner gebouwd is, daarna wordt besproken hoe analyse van feature vectors plaatsvindt en uiteindelijk de herkende zin wordt bepaald.

2.1.1. Het fonetisch schrift

In het fonetisch schrift is de uitspraak van een woord vastgelegd. Fonemen (betekenis-onderscheidende klanken) zijn de basis van de uitgesproken taal: alle woorden zijn akoestisch opgebouwd uit fonemen. Het fonetisch alfabet[1] lijkt veel op het gewone alfabet, zie voorbeeld tabel 2.2.

Voorbeeld.

encyclopedie in het Engels: encyclopedia [insajklapie:dia]

dit woord in het Nederlands: [ɛnsiklopedi]

Tabel 2.2 Tussen haken de fonetische transcriptie van het woorde “encyclopedie” in zowel Engels als Nederlands (bron: wikipedia, 30 dec ’05).

[1] Het fonetisch schrift is onder andere te vinden op http://nl.wikipedia.org/wiki/Fonetisch (30 dec ‘05). Hier zijn ook een uitgebreidere beschrijving en voorbeelden van fonetische woorden te vinden.

2.1.2. Opbouw van een herkenner

Een herkenner bestaat uit drie delen:

- Fonetisch lexicon
- Taalmodel
- Akoestisch model

Het lexicon bevat alle woorden die herkend moeten kunnen worden, met de fonetische transcriptie. Sommige woorden worden op verschillende manieren uitgesproken, bijvoorbeeld "Rotterdam": **/r O t @ d A m/** of **/r O t E r d A m/**. Voor deze woorden staan er meerdere transcripties vermeld in het lexicon.

Het taalmodel werkt als een soort grammatica. Het is een statistisch model, gebaseerd op *trigrammen*. Dit houdt in dat twee opeenvolgende woorden gebruikt worden om het derde woord te voorspellen. Het taalmodel geeft aan wat de kans op een bepaald woord is, gegeven de twee voorgaande woorden. Dit is niet hetzelfde als een grammatica, maar werkt zeer effectief in het voorspellen van correcte woordvolgorden.

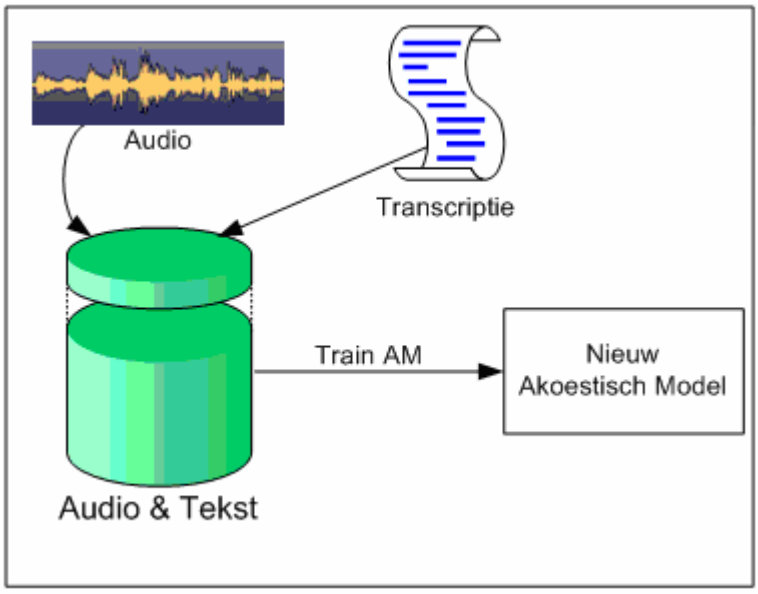
Het akoestisch model bestaat uit kansfuncties voor elk foneem. Zelfs een foneem kan elke keer anders uitgesproken worden, denk aan accent, man/vrouw, klemtonen en langzame sprekers. De kansfunctie van elk foneem (of: het foneem-model) geeft, gegeven een digitaal audiosignaal, de kans (akoestische likelihood) dat het foneem is uitgesproken. Er zijn ongeveer 50 fonemen, dus moeten er voor elk audiosignaal 50 kansen bepaald worden.

2.1.3. Training

Het lexicon dat gebruikt wordt door de herkenner *shout* voor herkenning in het broadcast-news-systeem, is een selectie van de woorden uit het Van Dale woordenboek met fonetische transcripties van elk woord.

Het taalmodel wordt getraind door aanbieden van teksten uit krantenartikelen. Dit corpus van meer dan 400 miljoen woorden wordt bijgehouden op de UT, en wordt gebruikt voor meerdere statistische onderzoeken. Het trainen bestaat uit het uitrekenen van alle woordcombinaties en de volgorde van woorden. Alle volgorden van drie elkaar opvolgende woorden (de trigrammen) wordt opgeslagen. De kans op een woord is dus afhankelijk van de twee voorgaande woorden. De score (voor een nieuwe herkenning) per trigram is de frequentie gedeeld door het totaal aantal verschillende trigrammen binnen de trainingsteksten.

De training van het akoestische model gebeurt door het aanbieden van gedecodeerde audiosignalen en de bijbehorende tekst. De audio wordt opgesplitst op de foneemgrenzen, waarna het juiste foneem uit de tekst hieraan gekoppeld wordt. Van elk foneem wordt een *foneem-model* getraind. Alle audiofragmenten van dat foneem worden gebruikt om de kansfunctie per foneem aan te passen. Alle foneem-modellen samen vormen het nieuwe akoestisch model.

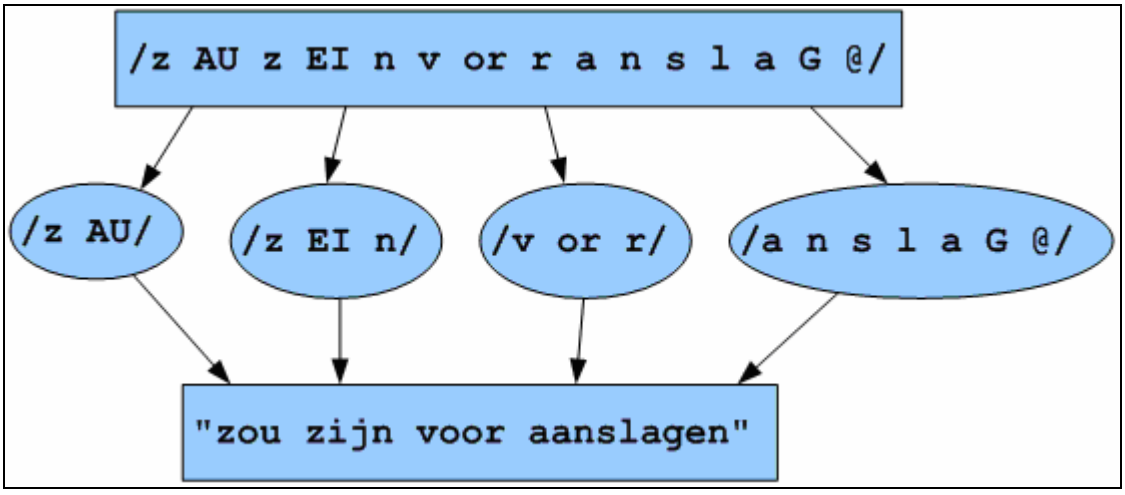


Tabel 2.3 De combinaties van audio en tekst (bij voorkeur transcriptie) wordt in een database bijgehouden. Al deze combinaties worden gebruikt om het (nieuwe) akoestisch model te trainen.

2.1.4. De herkenning

Van de herkenner wordt verwacht dat hele zinnen herkend worden. Niet alle zinnen zijn echter al eens uitgesproken, en het is niet te doen om modellen te bouwen voor hele zinnen. Het herkennen van hele woorden is zelfs erg moeilijk. Dit gebeurt alleen bij simpele systemen die een beperkt aantal voorgedefinieerde woorden herkennen. De moderne herkenners zijn getraind om fonemen te herkennen. Hier worden de basisstappen besproken die een herkenner moet nemen om van een audiosignaal tot volledige zinnen te komen. Herkenners als *shout* en *sonic* werken in de praktijk iets anders omdat ze geoptimaliseerd zijn voor snelheid.

De herkenner vergelijkt de geluidskenmerken met de foneem-modellen en zet vervolgens de digitale geluidsfragmenten om naar een serie losse fonemen. In tabel 2.4 is een serie losse fonemen te zien. In werkelijkheid is er een lijst met mogelijke fonemen op een bepaald tijdstip. Van elk foneem is de kans (likelihood) gegeven dat dit foneem ook de juiste is (de kans dat dit foneem juist “verstaan” is). De verzameling fonemen met kansen wordt de N-best-lijst genoemd.



Tabel 2.4 Voorbeeld fonetische transcriptie. De losse fonemen worden samengevoegd tot fonetische woorden, waarna er “echte” zinnen van gemaakt worden.

Vervolgens worden de foneemvolgorden vergeleken met de fonetische transcriptie van de in het lexicon aanwezige woorden. Op deze manier wordt van de losse fonemen woorden gemaakt. Hierdoor ontstaat een N-best lijst van woorden

De trigram-kansfunctie in het taalmodel wordt vervolgens gebruikt om van de lijst met mogelijke woorden, correcte zinnen te maken.

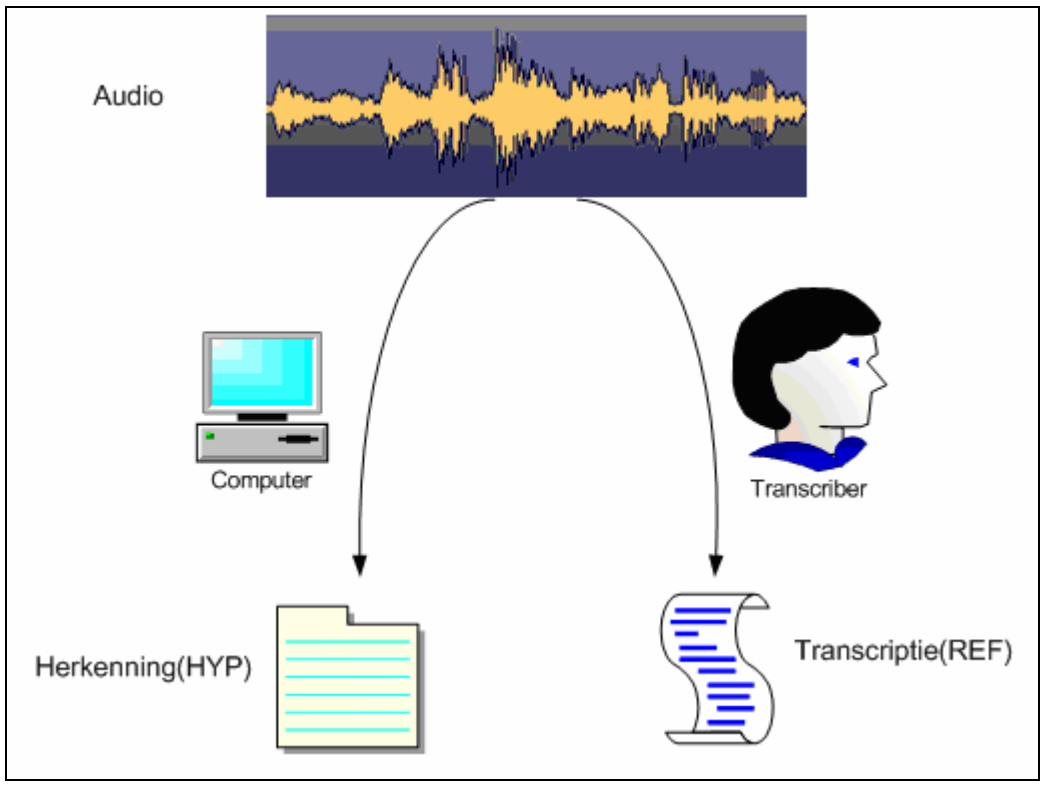
Meestal gaat het om een enorme hoeveelheid mogelijke foneemcombinaties en ook vrij grote hoeveelheid woordcombinaties. Er wordt in zowel het akoestisch model als het taalmodel gewerkt met kansen op een bepaalde uitspraak danwel woordvolgorde. De totaalscore voor een zin is een combinatie van akoestische score en taalmodel-score. De zin die uiteindelijk bovenaan staat in de N-best-lijst, en dus de hoogste kans heeft gekregen, wordt de output (herkenning).

2.2. *Verwerking van herkenning*

In dit onderzoek wordt gewerkt met herkenningen. Het is nuttig om de terugkerende (basis) begrippen als transcriptie, referentie, scoring en alignment hier al te lezen. Hier is geen onderzoek naar gedaan, maar ze worden wel telkens gebruikt.

2.2.1. Transcripties

De herkenning wordt vergeleken met een referentie om te bepalen hoe goed de herkenning is geweest. Om zeker te zijn dat de referentie van de herkenning goed is, moet het met de hand getranscribeerd woren. Hiervoor zijn programma's waar je als gebruiker na het horen van een stuk audio, kan intypen wat je verstaan hebt. Dit is de transcriptie.



Tabel 2.5 Aan de linkerkant het automatische herkennen, rechts herkennen door een persoon (transcriberen).

In tabel 2.5 is het idee van transcriptie en herkenning beschreven. De computer heeft de (gedigitaliseerde) audio als input, en levert een hypothesebestand op met daarin, naast allerlei andere informatie, de herkende tekst. De audio is vooraf automatisch in segmenten verdeeld. In

het beste geval zijn dat hele zinnen. De *transcriber* (oftewel annoteur) luistert naar dezelfde audio, en typt per segment in wat hij heeft verstaan. Per segment zijn nu een HYP en een REF beschikbaar.

2.2.2. Alignment

Bij alignment (of oplijning) worden twee zinnen met veel (maar niet allemaal) overeenkomstige woorden gekoppeld zodat duidelijk wordt welke woorden er overeenkomstig zijn, en welke in een van beide zinnen missen. Alignment wordt ook gebruikt bij DNA-onderzoek om twee molecuulketens te vergelijken.

Voor de toepassing binnen de spraakherkenning worden de hypothesezin (de herkenning) en de referentie (de transcriptie) gekoppeld, zoals bijvoorbeeld in tabel 2.6 is te zien. De woorden die niet gekoppeld kunnen worden aan een woorde van de andere zin, zijn in hoofdletters geschreven en worden foutgerekend. Bij een erg grote REF kunnen de woorden van de HYP vaak op veel plaatsen in de REF gemapt worden. Zoals in de figuur te zien is, zijn er zelfs in kleine zinnen soms meerdere oplijningen denkbaar.

```
REF:  dus de greep van DIE AUTOFABRIKANTEN is zo verschrikkelijk
HYP:  dus ** greep van DE  AUTOFABRIKANT   is zo verschrikkelijk
Eval:     D            S   S

REF:  dus ***** *** de GREEP         VAN DIE AUTOFABRIKANTEN is zo verschrikkelijk
HYP:  dus greep van de AUTOFABRIKANT *** *** *************** is zo verschrikkelijk
Eval:     I     I      S             D   D   D
```

Tabel 2.6 Figuur met twee oplijningen. De bovenste is de beste, de onderste een alternatieve oplijning, met matching op het woord “de”.

Een alignment-algoritme werkt volgens een dynamisch-programmeer-algoritme dat HYP en REF iteratief doorwerkt totdat beide zinnen geëindigd zijn(Huang, Hsu et al. 2003). Daartussen zoekt het voor alle woorden van de ene zin (in dit voorbeeld de HYP-zin), matches in de andere zin (in dit voorbeeld de REF). De voorwaarde is dat de volgorde van de woorden in het alignment dezelfde moet zijn als de volgorde in de HYP. De oplijning met de meeste matches is de output van het algoritme.

2.2.3. Scoring met Sclite

Sclite is een tool om twee teksten met elkaar te vergelijken. Het verzorgt alignment en de scoring. Sclite wordt voornamelijk gebruikt in de spraakherkenning, om hypothese (HYP) en referentie (REF) met elkaar te vergelijken. Het kan hele bestanden met HYP- en REF-zinnen verwerken en een score voor het hele bestand bepalen. De zinnen moeten allemaal gelabeld zijn zodat de juiste HYP-zinnen en REF-zinnen vergeleken kunnen worden. De woorden van de REF en HYP worden onder elkaar gezet, met de woorden die hetzelfde zijn precies onder elkaar (tabel 2.7). Daarna bepaalt Sclite welke woorden goed en fout zijn.

Insertions, deletions, substitutions

Er zijn drie typen fouten. Een fout wordt “insertion” genoemd als het woord in de HYP staat, maar niet in de REF. Een “deletion” is als het woord wel in de REF, maar niet in de HYP staat. Het is een “substitution” als op dezelfde plaats in HYP en REF een verschillend woord staat. Zie tabel 2.7 voor voorbeelden van deze fouten.

```
Scores: (#C #S #D #I) 8 2 1 1
REF:  DAN gaan ER in europa *** ANDERE regels gelden voor de autoverkoop
HYP:  DE  gaan ** in europa AAN DE     regels gelden voor de autoverkoop
Eval: S        D            I   S
```

Tabel 2.7 De Sclite-output. In de eerste rij de scores, in de volgende rijen de details. In REF en HYP zijn de fout-bestempelde woorden in hoofdletters. In de onderste rij zijn de fout-typen aangegeven (de evaluatie): S(ubstitution), D(eletion), I(nsertion).

Word-Error-Rate

Sclite levert een scorebestand op met de totaalscore voor de hele HYP, maar ook per zin het aantal insertions, deletions en substitutions.

$$WER = \frac{Insertions + Substitutions + Deletions}{Totaal\ aantal\ woorden\ REF}$$

Formule 2.1 De formule voor de Word-Error-Rate (WER). Insertions, substitutions en deletions zijn de verschillende fouten in herkenningen.

In dit afstudeeronderzoek staat de Word-Error-Rate (WER) centraal. In formule 2.1 is de formule hiervoor gegeven. De WER is een maat voor het percentage onjuist herkende woorden. In het ideale geval is dat 0%, maar goede systemen zitten slechts op 20%, afhankelijk van domein en toepassing. Hier is nog veel werk te verrichten dus.

3. De experimenten: statistiek

Omdat het in dit onderzoek om monitoring gaat (dus zonder tussenkomst van een mens), is het nooit helemaal zeker dat de prestatie zoals hier wordt gemeten (de voorspelling), gelijk is aan de daadwerkelijke prestatie van het systeem. De voorspelling moet zo veel mogelijk in de buurt liggen van de echte score. De interne confidence is een getal dat qua uiterlijk niet direct iets met word-error-rate (WER) te maken heeft (WER is in procenten, confidence is een getal tussen ongeveer -5 en +5). De externe confidence wordt uitgedrukt als een percentage, en heeft wel duidelijke gelijkenis met de “echte” WER en is daarom makkelijker te vergelijken.

Voor beide monitorings-methoden moet het verband tussen de waarde van de methode en de echte WER bepalen. Een maat voor de correlatie tussen twee variabelen is de correlatiecoëfficiënt.

Correlatiecoëfficiënt

De correlatiecoëfficiënt is een getal tussen -1 en 1 waarbij -1 een perfecte negatieve lineaire correlatie aanduidt en +1 een perfecte positieve lineaire correlatie. Een waarde van 0 betekent totaal geen lineaire correlatie(Basak 2006). Positieve correlatie is bijvoorbeeld: hoge WER volgens ondertiteling correleert met hoge WER volgens “echte” referentie. Negatieve correlatie is bijvoorbeeld een hogere interne confidence die samenhangt met een lagere WER. De formule voor de correlatiecoëfficiëntis te zien in formule 3.1.

$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma(X)\sigma(Y)}$$

Formule 3.1 De correlatiecoëfficiënt (Basak 2006).

De berekening is samengevat de volgende: per segment wordt de afwijking tot de gemiddelde X vergeleken met de afwijking tot de gemiddelde Y. Als in alle gevallen lineair verband is tussen afwijking tot X en afwijking tot Y, is de correlatiecoëfficiënt 1(of -1). Dit is zo als de coördinaten (x,y) een rechte lijn vormen in een assenstelsel (tenzij de lijn horizontaal of verticaal loopt). Hoe dichter de correlatiecoëfficiënt bij 0 ligt, hoe groter de afwijking is tot die lijn.

Beste correlatie

De experimenten die in de volgende hoofdstukken beschreven zijn, zijn er op gericht een zo hoog mogelijke (positieve danwel negatieve) correlatiecoëfficiënt te krijgen. Voor de voorgestelde aanpassingen zoals normalisatie van de interne confidence, of het extra ruim nemen van de referentie van de ondertiteling is telkens de correlatiecoëfficiëntbepaald. Daarmee zijn de beste monitoringsmethoden geselecteerd.

Bepalen van de correlatiecoëfficiënt

Het bepalen van de correlatiecoëfficiënt gebeurt natuurlijk niet met de hand. Daarvoor wordt het het softwarepakket SPSS(SPSS 2003) gebruikt. Het bepalen van de correlatiecoëfficiënt (met de zgn. *pearson*-methode) is een van de vele functies. Het is nodig om een overzicht te hebben met per zin, de waarden van deze twee variabelen naast elkaar. Voor het bepalen van de beste interne confidence bijvoorbeeld, wordt na elke variant op de normalisatie, een nieuw overzicht gemaakt met de nieuwe kandidaat-CM en de WER (zie 4.3, Experimenten). Aan de hand van dit overzicht wordt dan opnieuw de correlatiecoëfficiënt uitgerekend, en dit getal wordt vergeleken met andere kandidaat-confidence-measures.

4. Interne Confidence

In dit hoofdstuk wordt een bewerking van de interne akoestische likelihoods gezocht die voor elk segment een getal oplevert dat iets kan zeggen over de te verwachten WER in dat segment, en dus een mogelijkheid tot monitoring oplevert. Confidence-measures (al dan niet akoestisch) worden al veel gebruikt, daarom wordt eerst een uitgebreid literatuuronderzoek besproken. Daarna worden de meest bruikbare methoden uit de literatuur gebruikt voor een experiment op data van de herkenner *shout*. Daarin wordt geprobeerd de beste methode te vinden voor monitoring gebaseerd op akoestische confidences.

4.1. *Literatuur*

Likelihoods worden gebruikt door de herkenner om aan te geven welke fonemen het meest waarschijnlijk uitgesproken zijn, gegeven de gedigitaliseerde audiosignalen op een bepaald tijdstip (zie 2.1.4, “De herkenning”). Interne Confidences worden bepaald door bewerkingen op deze likelihoods van fonemen. Er zijn verschillende soorten bewerkingen, waarvan enkele worden toegelicht.

4.1.1. Verschillend gebruik herkenner

De eenvoudige spraakherkenningsystemen worden gebruikt voor isolated-word recognition. Dit is bijvoorbeeld bij dialoogsystemen zoals het treintijden-informatiesysteem. Daar stelt de computer de vragen, en moet de gebruiker antwoorden (plaatsnaam, ja/nee, vertrektijd etc). Daarnaast zijn er de command&control-systemen , waarin de gebruiker commando’s aan de computer kan geven, bijv (Damper, Tranchant et al. 1996) en (Couvreur, Boite et al. 2005). Hierbij hoeft slechts een beperkt aantal woorden herkend te worden. Er wordt daarom gewerkt met een beperkte grammatica en lexicon.

(Couvreur, Boite et al. 2005) en (Dolfing and Wendemuth 1998) bespreken dit soort systemen en het gebruik van confidencewaarden, om zowel de beste herkenning te selecteren uit de N-best lijst met beste herkenningen, als het accepteren danwel afwijzen (accept/reject) van een woord door instellen van een grenswaarde. In iets uitgebreidere systemen wordt keyword-recognition of –spotting gebruikt. Hierbij mag de gebruiker een hele zin gebruiken, en zoekt de herkenner de woorden die nodig zijn voor de input (Rose and Paul 1990).

Ingewikkelder wordt het bij *large-vocabulary-continuous-speech* herkenning (groot-vocabulaire, lopende spraak), waar alles herkend moet worden. Toepassingen zijn onder andere dicteersystemen (Thelen 1996), of broadcast-news herkenners zoals *shout,* (Huijbregts 2005), waar dit onderzoek over gaat.

4.1.2. Bewerking likelihoods

Likelihoods worden bewerkt afhankelijk van de toepassing en de mogelijkheden van het systeem. Via een inleiding over typen bewerkingen, worden de toepasbare bewerkingen voor broadcast-news-systemen besproken.

Waarom bewerkingen?

Bij isolated-word-recognition waarbij de gebruiker bijvoorbeeld heeft geantwoord in de “ja/nee” dialoog, heeft de herkenner een N-best-lijst (in dit geval met N=2) met “ja” en “nee”. Voor beide woorden zijn de likelihoods per foneem bekend. De eenvoudigste bewerking is nu om de likelihoods te vermenigvuldigen. De confidence-measure (CM) is het produkt van de afzonderlijke likelihoods. Vaak is het echter zo dat de herkenning bemoeilijkt wordt door een onduidelijke spreker of achtergrondgeluid. Ook kan de spreker meerdere woorden uitspreken (“dat klopt” of “nee bedankt”). De herkenner kan dit niet goed verwerken, maar zal toch voor “ja” en “nee”

waarden opleveren. Het is zelfs zo dat sommige fonemen beter herkend worden dan andere, waardoor ze standaard een hogere likelihood hebben. De confidencewaarden zijn er om een waarde te bepalen voor de herkenning, waardoor verschillende herkenningen correct vergelijkbaar zijn. Er kan een grenswaarde ingesteld worden om woorden te weigeren als ze niet hoog genoeg scoren. Het systeem kan dan bijvoorbeeld nogmaals naar het antwoord, of commando vragen.

Bij lopende spraak is het lexicon vele malen groter. Omdat het aantal woorden in een fragment spraak vrij groot is, is het niet altijd even duidelijk welke fonemen in welk woord thuishoren. Er wordt naast de akoestische likelihood ook een taalmodel-likelihood gebruikt om de meest logische woordvolgorde te bepalen. Hiermee wordt dus ook bepaald welke woorden gekozen worden (zie 2.1.4, “De herkenning”). Bewerking van akoestische likelihoods leiden er toe dat duidelijk kan worden welke woorden binnen een zin akoestisch gezien het best herkend zijn.

Bewerkingen: toepassing isolated-word-recognition

In (Gunawardana, Hon et al. 1998), gebaseerd op (Rose and Paul 1990), over keyword-spotting en isolated-word-recognition, wordt de basis-score per woord als hierboven berekend, het produkt van de losse foneem-likelihoods. De herkende woorden worden vergeleken met vooraf samengestelde lijsten met gemiddelde woordscores voor dat woord (berekend op dezelfde manier). Deze opgeslagen waarden worden ook wel “prior” waarden genoemd, wat “vooraf” betekent. De confidence wordt uiteindelijk bepaald door elk woord te delen met de opgeslagen “*prior”* waarde voor dat woord. Het trainen van elk woord is helaas voor large-vocabulary-systemen niet te doen.

De in (Dolfing and Wendemuth 1998), ook voor isolated-word-recognition, beschreven methode maakt gebruik van de N-best *posterior probability*. De woord-likelihood van het best herkende woord wordt gedeeld door de likelihood van de “2nd-best” herkenning, oftewel het tweede woord in de N-best lijst. De gedachte is dat er duidelijk verschil tussen de beste en 1-na beste herkenning moet zijn.

(Pinto and Sitaram 2005) gebruiken oa. een confidence die op foneem-lengte gebaseerd is. Per herkend foneem, wordt de lengte vergeleken met de gemiddelde lengte van dat foneem tijdens training. De mate van afwijking van dat gemiddelde duidt de kwaliteit van herkenning van dat foneem aan. De methode is goed bruikbaar per woord door foneem-likelihoods te vermenigvuldigen. Er wordt echter niets gezegd over lopende spraak.

Bewerkingen: lopende spraak

(Williams and Renals 1997) bespreken “prior-foneem-likelihoods” als akoestische confidence-measure. Hierbij wordt per woord de foneem-likelihood, of woord-likelihood gedeeld door de gemiddelde likelilhood voor die fonemen in de trainingsfase van de herkenner. Er wordt ook nog gesproken over lattice-rescoring, een methode die kijkt naar het aantal concurrerende woorden op een bepaald moment, en hieruit een confidence-waarde aan het herkende woord meegeeft. Deze methode is echter niet gebaseerd op akoestische kenmerken en is daarom niet onderzocht.

In (Mengosoglu and Ris 2005) worden twee confidence-measures gegeven. De eerste is op basis van prior-foneem-likelihoods informatie. Deze prior-methode is gelijk aan en zelfs afgeleid van (Williams and Renals 1997). De andere methode werkt met de maximum-foneem-likelihood van het best herkende foneem. In dit document zijn de beide voorgestelde confidencewaarden vergeleken met de fouttypen *ruis* (onduidelijke audio) en OOV (Out-of-Vocabulary, woorden die niet in het lexicon staan). De eerste methode zou vooral goed werken op audio met ruis, zoals audio over een onduidelijke kanaal als telefoonlijn, of met achtergrondmuziek. De tweede methode werkt vooral goed bij het detecteren van OOV’s, wat vaak neerkomt op nieuwe woorden, buitenlandse woorden of namen. Een lage score volgens een van beide methoden duidt een fout van deze respectievelijke categorieën aan.

Ook in (Shire 2001) wordt het gebruik van maximum-posterior probability's als confidence-measures besproken. Eenzelfde soort bewerking vindt daar plaats.

4.1.3. Normalisatie

Naast het bepalen van confidencewaarden door het delen door gemiddelde prior- en maximum posterior likelihoods, bespreken de bronnen normalisatiemethoden voor lopende spraak en keyword-spotting. Normaliseren houdt in dat een score gedeeld wordt door een lengtemaat, zodat zinnen, woorden of fonemen beter vergelijkbaar zijn. Zo staat bijvoorbeeld uitgelegd in (Couvreur, Boite et al. 2005): *"the confidence measure (..) ranges from 0 and -1 denoting high and low confidence, respectively."* . Normalisatie op tijd is nodig omdat *"Otherwise, longer utterances would always lead to lower confidence measures.*". Dit wordt verder uitgelegd in 4.2.3.

Naast normalisatie op tijd wordt er ook normalisatie op aantal fonemen en normalisatie op aantal alternatieve hypothesen gedaan. Voor het onderzoek wordt alleen met akoestische scores gewerkt en kunnen alternatieve hypothesen niet verwerkt worden.

4.1.4. Verder onderzoeken

De interne confidence-methoden die genoemd worden voor isolated-word-recognition zijn meestal alleen daarop (isolated-words) gericht en daardoor minder toepasbaar voor dit onderzoek. Er is wel een beeld ontstaan van de mogelijkheden. De methode met gemiddelde foneem-lengte(Pinto and Sitaram 2005) bijvoorbeeld is interessant, maar nog niet toegepast op lopende spraak, vandaar dat deze nu niet verder onderzocht wordt. Er wordt overigens wél rekening gehouden met de gemiddelde foneem-lengte bij normalisatie.

In de genoemde "lopende spraak"-bronnen zijn twee methoden genoemd: gebruik van average prior-foneem-likelihood (AVG) en gebruik van maximum posterior-foneem-likelihood (MAX) als uitgangspunt. In de eerste bron(Williams and Renals 1997) wordt AVG besproken, in (Shire 2001) de MAX-methode. In de derde, (Mengosoglu and Ris 2005) worden beide methoden besproken. De verwerking van likelihoods is hetzelfde, maar in de eerste en laatste bron wordt gezegd dat de methode goed werkt om te voorspellen of woorden goed of slecht herkend zijn, terwijl (Mengosoglu and Ris 2005) zeggen dat de methoden alleen voor een categorie fouten goed werken. Er is gekozen om zowel de AVG als MAX-methode te onderzoeken, en toch alleen te kijken naar de totale prestatie, dus fouten in het algemeen, niet per fout-type.

Toepassing van een akoestische confidence measure (CM) geeft de mogelijkheid om de kwaliteit van een herkenning uit te drukken in een getal. Dit getal staat voor het aantal fouten dat wordt verwacht in de herkenning. Hiermee kan de hele herkenning gemonitord worden.

Naast het onderzoeken van deze twee interne confidence-methoden worden ook verschillende normalisatiemethoden bekeken. Er worden in de verschillende bronnen twee normalisatiemethoden genoemd: delen door aantal tijdframes, en delen door aantal fonemen. De voor de hand liggende derde methode "delen door aantal woorden" is nog niet eerder besproken.

Waarschijnlijk is deze normalisatie minder onderzocht omdat de meeste systemen worden getest of ontwikkeld voor isolated-word-recognition. De lopende spraak-applicaties zijn gebaat bij het bepalen van fout-typen of gebruik van CM in N-best lijsten om de beste herkenning te bepalen. Hierbij is het slechts interessant om de herkenningen per woord te bekijken. Voor de toepassing in monitoring moet de hele zin bekeken worden, en zal ook normalisatie op aantal woorden plaatsvinden.

4.2. *Onderzoek*

In de vorige paragraaf zijn de verschillende interne confidence-methoden weergegeven. Er is aangegeven welke van toepassing zijn voor lopende spraak, en welke bruikbaar en haalbaar zijn

om een monitoring te kunnen bepalen. Hieronder worden de methoden verder uitgelegd, en de experimenten opgezet.

4.2.1. Confidences met behulp van “prior” informatie

Volgens (Mengosoglu and Ris 2005) is de confidence op basis van prior-informatie vooral nuttig om fouten met als oorzaak ruis of achtergrondgeluid te voorspellen. De data waarmee gewerkt wordt is ruisloze data van een nieuwsuitzending. Eigenlijk zou er ruis moeten worden toegevoegd aan de data om te controleren of deze methode voor broadcast-news werkt, maar er is voor gekozen om de methode te gebruiken als algemene fout-detector, en verder niet te kijken naar typen fouten. De redenering van (Mengosoglu and Ris 2005) is echter wél interessant. De methode gaat uit van vergelijking van de foneem-likelihood van de herkenning, met de gemiddelde likelihood van die foneem gedurende de trainingsfase. Als de score laag is, zijn de losse fonemen slechter herkend dan gemiddeld. Bij audio met ruis zullen de meeste fonemen dus slechter scoren dan gemiddeld. Een goede verklaring om aan te nemen dat de methode zal werken.

De confidence wordt als volgt bepaald: eerst worden de gemiddelde foneem-scores berekend. Deze worden in de trainings-fase van de herkenner bepaald (zie tabel 4.1).

$$\bar{P}(q_k) = \frac{1}{T} \sum_{t=1}^{T} P(q_k^t | X^t)$$

Formule 4.1 De gemiddelde score voor een bepaald foneem , waarbij T het aantal (zelfde) trainingsfonemen is. De likelihoods per foneem, P(q|X), worden opgeteld en gedeeld door T(Mengosoglu and Ris 2005).

De gemiddelden worden in de confidence-berekening gebruikt door de foneem-likelihoods te delen door de gemiddelde likelihood voor de betreffende fonemen. Van dat getal wordt het logaritme bepaald, waardoor het resultaat een getal is dat zowel positief (groter dan gemiddelde likelihood) of negatief (kleiner dan gemiddelde likelihood) kan zijn. De formule voor deze “average”-confidence, per woord, is te zien in tabel 4.2.

$$PPCM(W) = \frac{1}{N} \sum_{n=1}^{N} log(P(q_k^n | X^n) / \bar{P}(q_k))$$

Formule 4.2 De uitvoering van de confidence: de confidence van een woord (W) wordt bepaald door per foneem de likelihood, P(q|X), te delen door de gemiddelde score voor dat foneem. De log-scores van alle (N) fonemen worden bij elkaar opgeteld en gedeeld door het aantal fonemen in het woord(Mengosoglu and Ris 2005).

In de formule voor de confidence wordt gerekend met de som van het logaritme van de gemiddelde likelihoods. Dit is een applicatiegerichte techniek: optellen van logaritme is gelijk aan vermenigvuldigen van de gemiddelde likelhoods. Het is makkelijker te implementeren omdat er met kleinere getallen gewerkt wordt (likelihoods zijn soms in de orde van 10^{-100} of 10^{100}, dit is makkelijker weer te geven als -100 en 100).

Gemiddelde prior-score en shout

Voor dit onderzoek moeten dus gemiddelde prior-scores van alle fonemen vantevoren berekend worden. In *shout* worden alle scores direct logaritmisch opgeslagen. De gemiddelde prior-score per foneem is een gemiddelde prior log-score. Hierdoor is het de gemiddelde prior-score strict genomen niet helemaal correct. Vanwege het karakter van logaritmen kunnen grote uitschieters de “echte” gemiddelde waarde verstoren. Dit komt door de interne structuur van

shout, hier is op dit moment niets aan te doen. Omdat in de praktijk de prior-scores per foneem heel ver uit elkaar liggen, valt er echter prima mee te werken.

Verder zal het berekenen van de confidencewaarden grotendeels volgens de hierboven gegeven formules gaan. Er wordt een confidence per foneem berekend, een log-likelihood. Volgens formule 4.2 moeten de foneemscores opgeteld worden per woord, en gedeeld worden door het aantal fonemen in het woord. Dit is een vorm van normalisatie, deze worden nog verder onderzocht (zie ook 4.2.3).

4.2.2. Confidences met behulp van "posterior" informatie

Deze methode vergelijkt in een zin, in elk woord, de likelihood per foneem, met de maximum-waarde voor een foneem op dat moment. Het idee volgens (Mengosoglu and Ris 2005) is dat OOV's een lage likelihood hebben vergeleken met de maximum-likelihood op dat moment. Dat komt omdat het systeem een OOV niet kan herkennen en het akoestisch gezien best passende woord uit het lexicon invult. Dit woord zal een lage MAX-score krijgen. In formule 4.3 is de berekening van de confidence per woord gegeven.

$$RPCM(W) = \frac{1}{N}\sum_{n=1}^{N} log(P(q_k^n|X^n)/P(q_{best}^n|X^n))$$

Formule 4.3 Formule voor gebruik van posterior-max-likelihood. De confidence voor het herkende woord is de gemiddelde log-score per foneem. De log-score is de likelihood van het foneem gedeeld door de likelihood van het best herkende foneem op dat moment(Mengosoglu and Ris 2005).

De formule (de "max"-confidence) is vergelijkbaar met de formule voor gemiddelde foneem-score. Het bepalen van de beste foneem-scores wordt in *shout* opgelost door de audio nogmaals te herkennen, maar dan met een taalmodel dat alle foneem-combinaties toelaat, de zogenaamde "foneem-loop". Hierbij wordt eigenlijk alleen het akoestisch-model gebruikt waardoor de maximum-foneemscores bepaald worden. Het quotient van likelihood en maximum-likelihood zal een getal opleveren variërend van 0 tot 1 of logaritmisch minus oneindig tot 0.

4.2.3. Normalisatie van confidencewaarden

In de paragraaf 4.1, Literatuur werd al aangegven dat normalisatie nodig is om woorden correct te kunnen vergelijken. Er is ook geconcludeerd dat er verschillende normalisaties worden gebruikt, en de normalisatie op basis van aantal woorden in een zin, extra onderzocht moet worden. De normalisaties voor toepassing van monitoring zullen niet per woord getest worden, maar per hele zin. Dit is omdat de monitoring per zin (segment) geschiedt.

Waarom normalisatie?

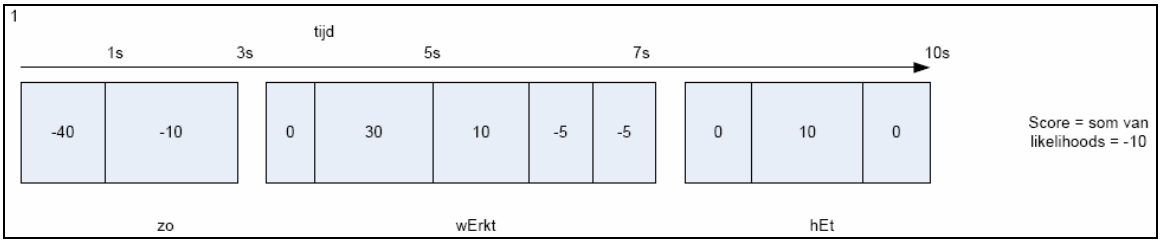
In (Couvreur, Boite et al. 2005) werd al aangegeven dat normalisatie nodig is omdat anders langere woorden altijd een lagere score hebben dan korte woorden, door het vermenigvuldigen van likelihood (kansen). Als de gebruiker bijvoorbeeld een commando van 100 tijdframes (1 seconde) lang geeft, met gemiddelde likelihood van 0.8, zal de score 0.8^{100} zijn, wat een CM van ongeveer $2 * 10^{-10}$ heeft, oftewel een log-score van -10. Een langer commando van dezelfde kwaliteit, bijvoorbeeld 2 seconden, zal een CM van $0.8^{200} = 4*10^{-20}$ hebben, oftewel -20. Als het systeem voor acceptatie/weigering van een commando een grenswaarde heeft van -15, zal het ene commando wél geaccepteerd worden en het andere niet. De voorgestelde oplossing: totaalscore delen door de lengte van het commando levert voor beiden -1 op, waarmee aangetoond is dat normalisatie de woorden beter vergelijkbaar maakt.

Zoals in 4.1.4 is aangekondigd, zal het onderzoek zich richten op segmenten: hele zinnen, of soms zelfs meerdere zinnen. De in de literatuur aangegeven methoden moeten daarvoor opnieuw

onderzocht worden. Om toepassing binnen zinnen te verduidelijken zijn in dit hoofdstuk de verschillende normalisaties uitgelegd.

Hoe werkt normalisatie?

De werking van de verschillende normalisatiestrategieën wordt hieronder uitgelegd. De uitgangspositie is een segment van een bepaalde tijd T, een aantal woorden W en een aantal fonemen N. Alle fonemen hebben een bepaald score (log-likelihood). Als er geen normalisatie gebruikt wordt, is de score voor het segment de som van de log-scores (wat dus eigenlijk het produkt van de likelihoods is, oftewel de basis-confidence). Deze situatie is geillustreerd in tabel 4.1.



Tabel 4.1 De score is de som van de afzonderlijke likelihoods, zonder normalisatie.

Dit is de basis. De berekening is als in formule 4.4.

$$Basis\_CM = \sum_{i=1toN} L_i$$

Formule 4.4 De basis-confidence is de som van de Log-likelihoods per foneem.

Er is uitgelegd dat langere zinnen van dezelfde kwaliteit slechtere of juist betere waarden krijgen, afhankelijk van of de basis negatief of positief is.

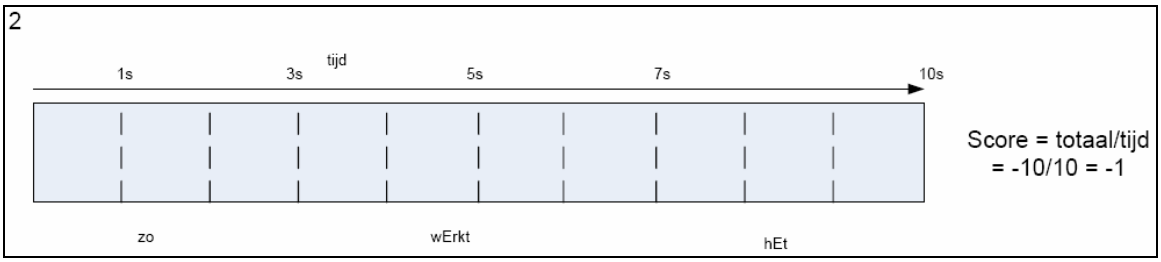
Tijd-normalisatie

De foneem-scores van een zin of segment worden bij elkaar opgeteld en gedeeld door de tijd dat dat hele segment duurt. Formule:

$$Time\_norm\_CM = \frac{\sum_{i=1toN} L_i}{T}$$

Formule 4.5 De confidence is de totaalscore (som van de foneem-likelihoods) gedeeld door de totaaltijd dat het segment duurt N = het aantal fonemen in het segment, L_i is de foneemscore van foneem i, en T is de tijd dat het segment duurt.

In het voorbeeld van figuur tabel 4.1 zou tijdnormalisatie er uitzien als in tabel 4.2.



Tabel 4.2 Het segment wordt als een geheel gezien, slechts bestaande uit een totaalscore en een totaaltijd. De stippellijnen geven de seconden aan.

Voor elk segment wordt dit gedaan, zodat de segmenten gezien kunnen worden alsof ze van de zelfde lengte zijn, en correcter vergeleken kunnen worden.

Foneem-normalisatie

De foneem-scores van een segment worden bij elkaar opgeteld, en gedeeld door het aantal fonemen in dat segment. De formule voor de CM is te zien in formule 4.6.

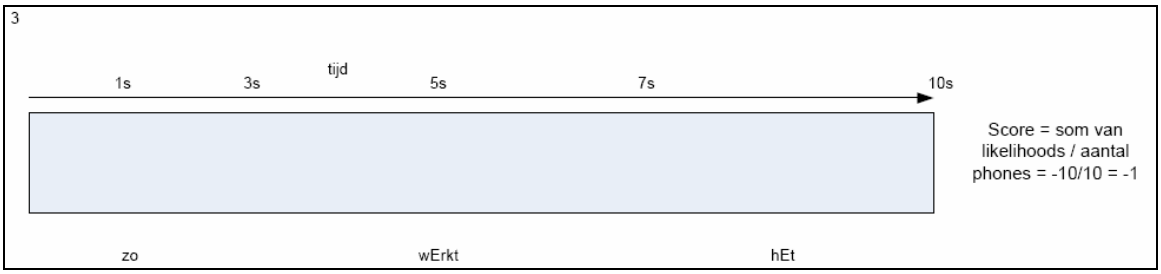
$$Phone\_norm\_CM = \frac{\sum_{i=1toN} L_i}{N}$$

Formule 4.6 De confidence is de som van de foneemscores, gedeeld door het aantal fonemen in het segment.

Hiermee wordt een segment gezien als een geheel van 1 foneem (zie tabel 4.3). Het verschil tussen segmenten is nu de **gemiddelde** foneemlengte. Hoewel er dus niet echt op tijd is genormaliseerd, kan deze normalisatie toch gezien kan worden als een lengte-normalisatie omdat langere zinnen ook vaak meer fonemen hebben.

Daarnaast worden door foneem-normalisatie korte fonemen belangrijker beschouwd dan lange fonemen. Extreem korte fonemen treden op als een foneem erg slecht herkend wordt. Het akoestisch signaal krijgt dan een slechte score volgens de foneem-modellen. De herkenner lost dit op door de fonemen erg kort te maken, waardoor de totaalscore van een woord niet extreem slecht wordt. De slechte foneem wordt als het ware “weggestopt”. Door normalisatie wordt de totaalscore per zin lager, waarmee zo’n korte foneem extra benadrukt wordt.

In tabel 4.3 is de schematische weergave te zien na toepassing van foneem-normalisatie.



Tabel 4.3 Het segment wordt gezien als geheel van 1 foneem.

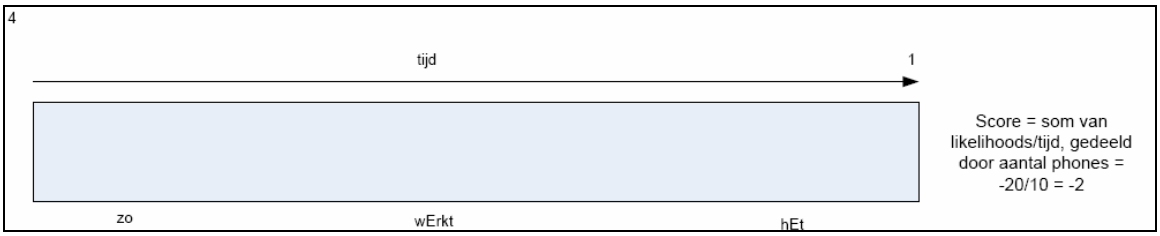
Combinatie van beiden

Hierbij wordt per foneem genormaliseerd op de tijd die dat foneem duurt, en daarna per segment genormaliseerd op het totaal aantal fonemen (formule 4.7).

$$PHT\_norm\_CM = \frac{\sum_{i=1toN} \frac{L_i}{T_i}}{N}$$

Formule 4.7 De normalisatie op tijd, en daarna op aantal fonemen., met L_i de foneemscore, T_i de tijd dat foneem i duurt, en N het totaal aantal fonemen.

Deze dubbele normalisatie zorgt er voor dat alle segmenten gezien worden als 1 foneem met lengte 1 (tabel 4.4).



Tabel 4.4 Het segment wordt gezien als een groot foneem, van tijd 1.

Dit is de laatste normalisatie uit de literatuur die besproken wordt. Er wordt geen rekening gehouden met het aantal woorden per segment. Dit is het enige verschil tussen de segmenten. Het zal blijken of het gemiddeld aantal woorden van invloed is op de WER van het segment.

Woord-normalisatie

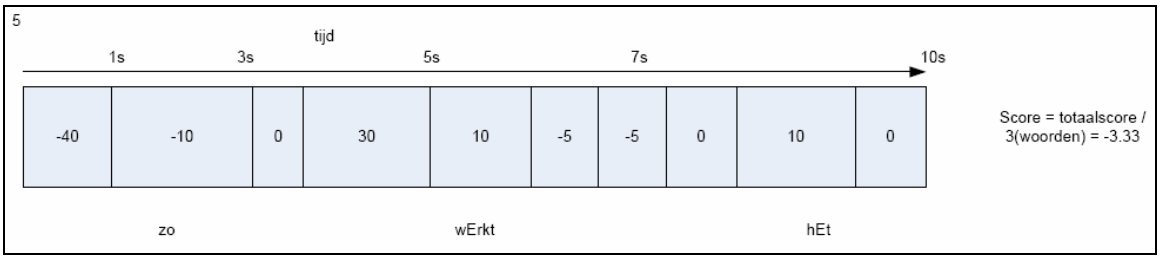
Bij woord-normalisatie wordt de score van een segment gedeeld door het aantal woorden in de zin (formule 4.8).

$$Woord\_norm\_CM = \frac{\sum_{i=1toN} L_i}{W}$$

Formule 4.8, de confidence voor basis-woord-normalisatie. De som van alle foneemscores wordt gedeeld door het aantal woorden in de zin.

Het idee achter deze normalisatiemethode is dat het uit zou kunnen maken uit hoeveel woorden een segment bestaat, net zoals er in de foneem-normalisatiemethode van wordt uitgegaan dat het aantal fonemen in een woord verschil maakt. Deze basis-woordnormalisatie is echter slechts een aanduiding van tijd. Het is logisch dat een segment met meer woorden langer duurt. Hiermee is deze vorm eigenlijk een variant van de tijdnormalisatie.

In tabel 4.5 is de grafische weergave van woord-normalisatie te zien.



Tabel 4.5 het segment wordt gezien als een woord, waarin de oorspronkelijke tijd en fonemen blijven bestaan.

De vergelijking met andere segmenten is anders dan de in de literatuur genoemde methoden, er wordt niet gekeken naar lengte in tijd of aantal fonemen. Bij de andere normalisaties worden de zinnen vergeleken zonder te kijken naar het aantal woorden. Deze methode vergelijkt door alleen te kijken naar het aantal woorden. Er wordt niet gekeken naar het aantal fonemen of de duur van de zin. Het is tevens interessant om naast aantal woorden, hier tóch ook naar te kijken. Daarom zijn ook combinaties van voorgaande drie methoden en woord-normalisatie onderzocht. Naast tijd, fonemen en beiden bestaan de normalisaties dan uit:

- woorden (formule 4.8)
- tijd en woorden (formule 4.9)
- fonemen en woorden (formule 4.10)
- tijd, fonemen en woorden (formule 4.11)

Voor het toepassen van de hier genoemde normalisaties wordt per woord de totale foneem-score en totale tijd bijgehouden. Normalisatie op tijd en/of foneem gebeurt dan per woord. In formulevorm:

$$TW\_norm\_CM = \frac{\sum_{j=1\,to\,W} \frac{\sum_{i=1\,to\,N_{wj}} L_i}{T_{wj}}}{W}$$

Formule 4.9 Normalisatie op tijd en aantal woorden.

$$PW\_norm\_CM = \frac{\sum_{j=1\,to\,W} \frac{\sum_{i=1\,to\,N_{wj}} L_i}{N_{wj}}}{W}$$

Formule 4.10 Normalisatie op aantal fonemen en aantal woorden.

$$ALL\_norm\_CM = \frac{\sum_{j=1\,to\,W} \frac{\sum_{i=1\,to\,N_{wj}} \frac{L_i}{T_i}}{N_{wj}}}{W}$$

Formule 4.11 Normalisatie op alle drie lengtematen.

Met nogmaals: W is het aantal woorden, L_i de likelihood van foneem i, T_i de lengte in tijd van foneem i, N_{wj} het aantal fonemen in woord j. De laatste categorie normaliseert op tijd, fonemen en woorden, en beschouwt alle segmenten als bestond het uit 1 woord van 1 foneem, van 1 tijdsframe.

4.2.4. Experiment-opzet

Er is beschouwd wat de interessante onderdelen zijn qua verschillende interne confidencewaarden en normalisaties. De in de literatuur gevonden methoden voor interne confidencewaarden en normalisaties zullen worden gecontroleerd door toepassing binnen het systeem *shout*, maar ook de normalisatie op basis van aantal woorden per zin, zal worden getest. Er zullen daarom 2 (confidence-typen) * 8 (normalisaties) = 16 experimenten gedaan worden. De categorie met de hoogste correlatie-coefficient tussen interne confidence en WER (zie hoofdstuk 3, *De experimenten: statistiek*) wordt aanbevolen voor de monitoring-tool en zal gebruikt worden bij het experiment adaptatie (zie hoofdstuk 6, Training).

In bijlage A zijn alle scripts beschreven die gebruikt zijn voor de verschillende methoden.

4.3. *Experimenten*

In dit hoofdstuk worden de aangekondigde experimenten over de beste confidence-measure besproken. Bij de experimenten staat de corelatiecoefficiënt centraal. De zestien categorieën zijn in de volgende figuur (tabel 4.6) weergegeven.

Correlatiecoefficienten tussen geteste CM en WER			
		Confidence	
Normalisatie		Avg-likelihood-methode	Max-likelihood-methode
methode	1. geen		
	2. tijd		
	3. phone		
	4. woord		
	5. tijd&phone		
	6. tijd&woord		
	7. phone&woord		
	8. alles		

Tabel 4.6 blanco tabel met daarin de te onderzoeken categorieën. 8 normalisatiemethoden, 2 typen confidence-methoden.

4.3.1. Inleiding

Om confidencewaarden te bereken volgens de formules van 4.2 is het nodig om van een herkenning, per foneem de likelihoods te bepalen. Verder moeten de zinnen per woord doorlopen worden zodat ook normalisatie per woord (4 versies) gedaan kan worden. De uitvoerbestanden van Shout bevatten de interne likelihoods per foneem, zie tabel 4.7. Daarnaast zijn de fonemen per woord gescheiden.

```
ID: 20020112s000-000
n   |  0   2 | -47.45464     -24.3927
O   |  3  23 | -32.57334     -10.28545
s   | 24  43 | -46.61258     -27.84507

Z   | 49  59 | 104.84528     -13.08084
u   | 60  66 |  20.22079      -7.36844
r   | 67  71 |-100.06351     -22.84158
n   | 72  79 |-129.90381     -28.0874
a   | 80 125 |-153.82104     -12.68362
l   |126 128 | -21.54413       3.16992
```

Tabel 4.7 Een deel van een *shout*-output-bestand. Deze zin bestaat uit de woorden "NOS, Journaal", in fonetisch schrift. Naast de afzonderlijke fonemen zijn de tijdframes getoond (tussen pipes), de likelihood van het foneem, en de max-likelihood (van het best herkende foneem op dat moment).

Dit is de structuur van de output van de herkenner. Eerst worden de tijdframes gegeven. Het eerste foneem ("n") duurt van tijdframe 0 t/m 2. Een frame duurt 10ms, dus de eerste klank "n" duurde 30ms. Daarna volgen de likelihoods. De eerste kolom is de foneem-score voor de "n". De tweede kolom is score van het best herkende foneem op dat moment. De waarde in de eerste kolom kan dus in theorie nooit groter zijn dan in de tweede. In de praktijk kan dit door afrondfouten af en toe dit wel voorkomen. In de figuur is echter te zien dat bij twee fonemen de eerste waarde vele malen groter is dan de tweede. Dit is alleen te verklaren als een foutje in de herkenner. Helaas is dit tot op dit moment niet opgelost en wordt er gewoon met de getallen verder gerekend.

Een uitvoerbestand begint met de gemiddelde scores van alle fonemen. Deze gemiddelden zijn bepaald op basis van andere herkenningen. Voor elke klank is dus de gemiddelde score bepaald. Dit is voor het bepalen van de gemiddelde foneem-score, voor de eerste confidence-methode.

```
.
.
.
# Z   (407) -  -56.864019
# a (40036) -  -0.295934
# b (20900) -  -62.268979
# d (68588) -  -63.360178
# e (58521) -  -15.025976
.
.
etc
```

Tabel 4.8 begin van elk uitvoerbestand van *shout*, de gemiddelde likelihoods voor alle fonemen.

Bepalen score

Het bepalen van de confidencewaarden gaat als volgt. Om te beginnen moet per foneem de likelihood worden bepaald. Bijvoorbeeld van het foneem 'a', uit tabel 4.7. De 2e kolom van rechts bevat de basis-likelihood. Voor de "average"-methode, wordt de "prior"-informatie uit tabel 4.8 gehaald. De average-score voor het foneem 'a' is -153.8 - -0.30 = -154.1. Dit betekent dat het foneem 'a' in deze zin eenveel lagere akoestische score heeft dan in de trainingsset.

Voor het bepalen van de "max"-score voor 'a' wordt de laatste kolom uit tabel 4.7 gelezen. De max-score is -153.8 – -12.7 = -141.1. Dit betekent dat 'a' verre van perfect herkend is.

Voor beide confidencewaarden geldt dat er in de formules gedeeld wordt, en hier afgetrokken. Dit komt omdat er in het deel-gedeelte van de formules niet op logaritmische schaal wordt gewerkt, maar de likelihoods die in *shout* worden gegenereerd zijn wel op logaritmische schaal. Dit geldt echter ook voor de max-waarden (van de foneem-loop) en de average-waarden van de trainingsset (prior informatie). De confidencewaarden worden verder wel op dezelfde manier berekend.

Normalisatiemethoden

Het bepalen van de 16 confidencewaarden gebeurt door per ID, per woord, per foneem de score in te lezen en op te slaan, plus het aantal tijdframes. De totaalscore is de som van de losse foneem-scores. Het normaliseren op tijd gebeurt door de totaaltijd op te slaan en de totaalscore daardoor te delen. Normaliseren op aantal fonemen gebeurt door de fonemen te tellen en de totaalscore daardoor te delen. De combinatie gebeurt door elke foneemscore direct te delen door de tijd van dat foneem en de som van die scores te delen door het aantal fonemen. Alle scores worden verder berekend volgens de formules van 4.2.

Gegevens

Voor het experiment wordt gewerkt met het 8-uur-journaal van 12 jan '02, van 20 minuten, met als bruikbare data 116 zinnen met foneem-scores en transcriptie.

De prestatie wordt gemeten aan de hand van de correlatiecoëfficiënt van de segmenten met hun WER (zoals eerder gezegd). Omdat er transcriptie beschikbaar is, is van elk segment al vooraf de WER bepaald. Dit is gedaan met Sclite (uitgelegd in 2.2.2).

4.3.2. Uitvoering experimenten

De experimenten zijn uitgevoerd in volgorde van introductie. Het doorlopen van een bestand gaat per segment, per foneem. De bewerking van een foneem gebeurt zodra de likelihoods zijn ingelezen. Hierna worden direct de score per foneem volgens de AVG-methode en de MAX-

methode berekend. Na het inlezen van het hele segment wordt de gewenste normalisatiemethode uitgevoerd.

Er wordt van twee experimenten de hele confidence-berekening getoond. Omdat de berekening volgens de formules is, is het overbodig om voor alle 16 methoden de berekeningen te tonen. De getoonde experimenten laten de methoden “zonder normalisatie” en “normalisatie op tijd” zien.

Eerste experiment: geen normalisatie

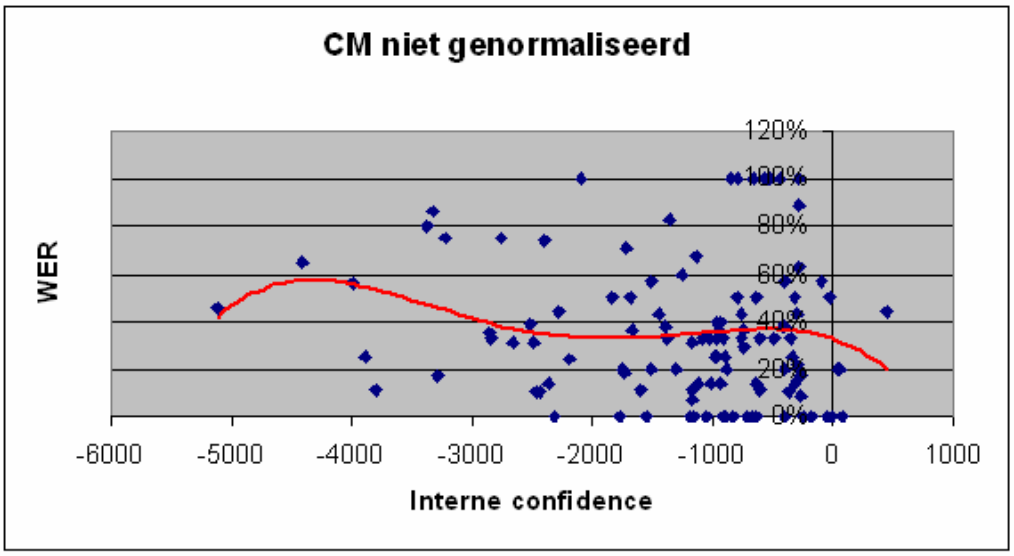
Hierbij worden voor alle zinnen, de scores per foneem bij elkaar opgeteld. De CM’s per zin lopen daardoor nogal uit elkaar (nogmaals: slecht scorende fonemen halen het getal erg naar beneden).

Het voorbeeld van de zin “nos journaal” is hier genomen omdat het kort is en al eerder geintroduceerd is. Eerst wordt de average-confidence bepaald.

Phone	Duur	log-likelihood	Avg-LL	Max-LL	AVG	MAX
n	3	-47.45	-57.87	-24.39	10.41	-23.06
O	21	-32.57	-48.19	-10.29	15.61	-22.29
s	20	-46.61	-74.38	-27.85	27.77	-18.77
Z	11	104.85	-56.86	-13.08	161.71	117.93
u	7	20.22	-57.52	-7.37	77.74	27.59
r	5	-100.06	-63.10	-22.84	-36.97	-77.22
n	8	-129.90	-57.87	-28.09	-72.04	-101.82
a	46	-153.82	-0.30	-12.68	-153.53	-141.14
l	3	-21.54	-61.40	3.17	39.86	-24.71

Tabel 4.9 De gegevens waarmee de average-confidence bepaald wordt. Uitspraak in fonemen, hun lengte, de huidige foneemscore, de gemiddelde scores voor dat foneem, de maximumscore op dat moment, en de AVG en MAX-waarden.

De totaalscore van dit segment is, afhankelijk van de methode, de som van kolom “AVG” danwel “MAX” en is de waarde 70 respectievelijk -263. Met een word-error-rate van 0% is dit segment in grafiek 4.1 te vinden op het punt (70, 0), en in grafiek 4.2 op het punt (263, 0) In dit figuur staan coördinaten van alle segmenten van dit journaal.



Grafiek 4.1 Voor alle waarden van de interne confidence is de WER gegeven. In het rood het interpolatiepolynoom.

In dit figuur is te zien dat de waarde van de WER nogal uiteenloopt. Uit de blauwe punten lijkt moeilijk een verband te bepalen tussen Interne confidence en WER. De rode lijn (het interpolatiepolynoom) geeft echter de gemiddelde waarde weer. Op basis van de licht dalende rode lijn kan vermoed worden dat een hogere CM in verband staat met een lagere WER. Het is echter een zeer licht dalende lijn (gemiddeld 20% WER op 5000 CM-punten). Bovendien is er de mogelijkheid om een correlatiecoëfficiënt te bepalen, waarbij de verschillende methoden tot een getal leiden. Dat is makkelijker vergelijken.

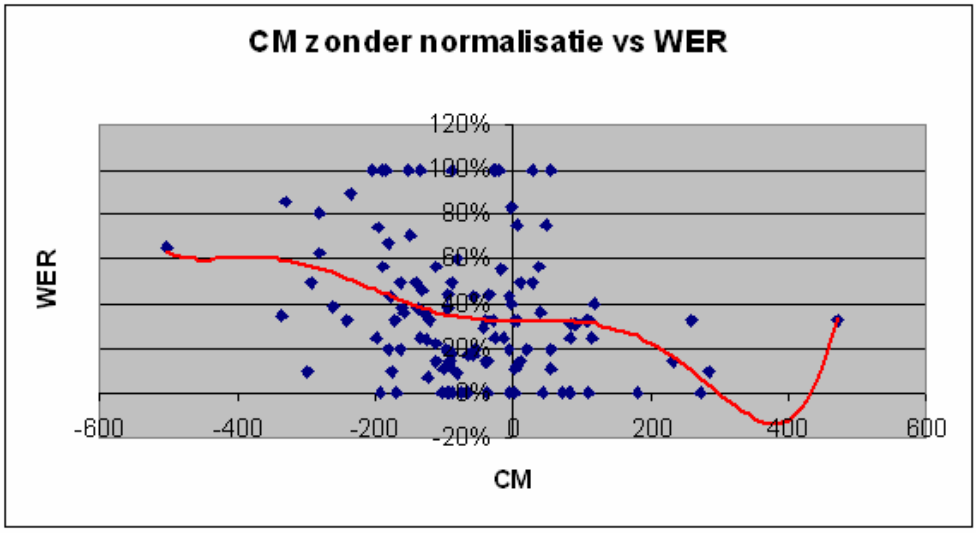
Correlations

		hrReal	Conf
hrReal	Pearson Correlation	1	-.094
	Sig. (2-tailed)		.317
	N	116	116
Conf	Pearson Correlation	-.094	1
	Sig. (2-tailed)	.317	
	N	116	116

Tabel 4.10 De SPSS-output voor de correlatiecoëfficiënt tussen alle confidencewaarden en WERs van het jouranaal van 12-01-02. Rechtsboven de correlatie tussen "hrReal "en "Conf", oftewel WER vs CM.

De correlatiecoëfficiënt is -0.094. Dit geeft aan dat er over het gehele domein van de CM, een licht negatief verband is tussen CM en WER. Een hogere CM staat dus inderdaad in verband met een lagere WER, al is het een klein verband. Voor de "max"-confidencewaarden is op eenzelfde soort manier met de gegevens omgegaan.

De grafiek van de max-confidence tegenover de WER is weergegeven in grafiek 4.2.



Grafiek 4.2 Max-confidence zonder normalisatie, uitgezet tegen de WER. De rode lijn is het interpolatiepolynoom dat de gemiddelde waarden van de punten aangeeft.

Er is al verteld dat er waarschijnlijk een foutje in de herkenner zit waardoor de MAX-waarden structureel te hoog zijn. Dit heeft tot gevolg dat in de grafieken veel positieve confidence-waarden staan, zelfs nog meer dan bij de AVG-methode. Dit zou niet mogelijk moeten zijn, maar er is toch verder gewerkt met deze waarden.

Correlations

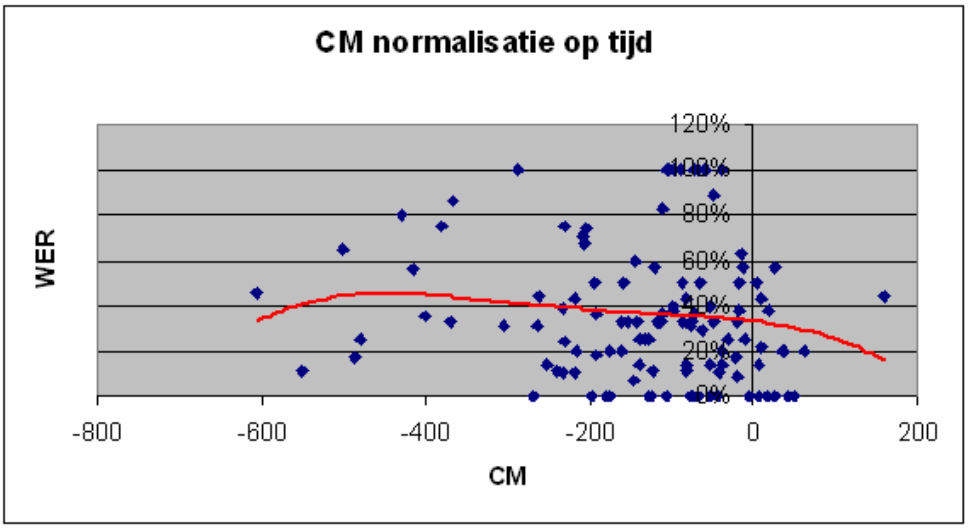
		hrReal	Conf
hrReal	Pearson Correlation	1	-.264**
	Sig. (2-tailed)		.004
	N	116	116
Conf	Pearson Correlation	-.264**	1
	Sig. (2-tailed)	.004	
	N	116	116

Tabel 4.11 De SPSS-output voor de max-confidencewaarden. De correlatiecoëfficiënt staat rechtsboven in de tabel en is -0.264. De tweede rij geeft een andere correlatie-methode aan, de derde het aantal zinnen (116).

Zoals te zien is in tabel 4.11 is de correlatiecoëfficiënt een stuk lager dan bij de average-CM. Dit duidt op een veel duidelijkere negatieve correlatie tussen de max-categorie en de WER dan dat de average-categorie heeft.

Andere normalisaties

Van alle normalisatiemethoden, voor beide confidence-methoden zijn grafieken gemaakt. Van de normalisatiemethode “tijd” zijn de waarden in grafiek 4.3 te vinden. Er is een groot verschil met grafiek 4.1 qua confidence-waarden. De waarden in de grafiek liggen veel dichter bij 0. De interpolatielijn is echter nog steeds licht dalend.



Grafiek 4.3 De confdiences van de segmenten uitgezet tegen de WER. De confidencewaarden zijn bepaald volgens de normalisatiemethode: tijd, en de average-confidence-methode.

Van alle andere categorieën zijn de grafieken te vinden in bijlage B.

4.3.3. Resultaten: correlaties

De correlatiecoëfficiënt van alle normalisatiemethoden en confidence-methoden zijn gegeven in tabel 4.12, de eerste rij was al besproken.

Correlatiecoefficienten tussen geteste CM en WER			
		Confidence	
Normalisatie		Avg-likelihood-methode	Max-likelihood-methode
methode	1. geen	-0.094	-0.264
	2. tijd	-0.332	-0.355
	3. phone	-0.335	-0.354
	4. woord	-0.323	-0.319
	5. tijd&phone	-0.362	-0.413
	6. tijd&woord	-0.341	-0.260
	7. phone&woord	-0.272	-0.272
	6. alles	-0.288	-0.372

Tabel 4.12 overzicht van correlatiecoëfficiënten. Links de normalisatiemethoden, boven de confidence-methoden. Van de best presterende categorie is de rij omlijnd.

De getallen in de tabel geven aan dat alle categorieën een negatieve correlatie met de WER hebben. Het doel van deze experimenten was om te bepalen wat de beste confidence-methode is, gecombineerd met de beste normalisatiemethode, om de WER te voorspellen.

Allereerst valt op dat de CM voor de average-methode, zonder normalisatie, een correlatiecoëfficiënt van net onder 0 heeft. Dit is onverwacht. Zoals in hoofdstuk 3 aangegeven is, zou te verwachten zijn dat de correlatiecoëfficient liefst lager dan -0.1 is om te kunnen zeggen dat er een duidelijk verband is. De max-methode laat echter wel degelijk een duidelijk verband zien. Dit is opmerkelijk, het lijkt dat de niet-genormaliseerde versie van de average-methode bijna niets zegt over de WER. De categorieën waarbij wél normalisatie is toegepast, presteren beter. Bij de max-methode hebben alle categorieën een duidelijke correlatie met de WER. Het zou kunnen dat dit komt omdat bij de MAX-methode de fonemen die vergeleken worden dezelfde lengte hebben.

De AVG-methode vergelijkt fonemen van verschillende lengte, wellicht dat er daarom weinig verband met de WER te vinden is.

De beste categorie is de confidence-methode max, oftewel gebruik van posterior likelihoods, in combinatie met de normalisatie-methode “tijd&foneem”, oftewel eerst de foneem-scores delen door de foneem-lengte in tijd, en daarna de segmenten delen door het aantal fonemen per segment.

Welke confidence-methode?

Uit de resultaten is af te leiden dat de MAX-confidence in de meeste normalisatiecategorieën beter werkt dan de AVG-confidence. Bij beide confidence-methoden levert de normalisatie-methode “tijd&foneem” de beste correlatie. Daarin is het verschil tussen beide confidence-methoden niet zo heel groot, maar toch groot genoeg om te concluderen dat “tijd&foneem” de beste is.

Volgens (Mengosoglu and Ris 2005) werkt de average-methode goed op data met ruis, en de max-methode op data met veel OOV’s. De resultaten laten zien dat ze allebei wel aardig werken, de max-methode iets beter. Het was niet mogelijk om experimenten te doen met geprepareerde data met ruis of veel OOV’s. Wellicht dat de resultaten dan anders zouden zijn. Tevens is het mogelijk dat ze juist op hun eigen gebied goed presteren, en dat een combinatie van beide methoden nóg beter presteert. Deze combinatie is verder niet uitgewerkt, dus de max-methode is op dit moment de beste.

Woord-normalisatie

Hoewel gebruik van woord-normalisatie (zie tabel 4.12, categorie 4) goede resultaten oplevert, bijna vergelijkbaar met tijd en aantal fonemen, is het toch niet de beste keus geworden. Het is mogelijk om op aantal woorden te normaliseren, omdat dit een zekere aanduiding van lengte van het segment aanduidt.

4.4. *Conclusies*

Monitoring op basis van interne likelihoods lijkt goed mogelijk. In de literatuur worden likelihoods vooral gebruikt om acceptatie/afwijzing van losse woorden toe te passen. Een andere bron voor lopende spraak geeft methoden die fouttypen kunnen voorspellen, maar nu is ook aangetoond dat in een heel journaal met segmenten van ongelijke grootte, likelihoods een indicatie kunnen geven van welk percentage van de zin fout is.

In 4.2.1, “Gemiddelde prior-score en shout”, is aangegeven dat de gemiddelde prior-foneemscores niet helemaal goed bepaald kunnen worden. Het is mogelijk dat als dit wel kan, de “average”-methode beter werkt dan dat het nu doet. Op dit moment is de methode volgens maximum posterior-foneem-likelihoods, de beste. Hier wordt per foneem de likelihood gedeeld door de likelihood van het best scorende foneem op dat moment (*de relatieve score*). Deze waarden moeten genormaliseerd worden door de score per foneem te delen door het aantal tijdframes dat het foneem duurt, en daarna de totaalscore van fonemen te delen door het aantal fonemen. Hiermee heeft de interne confidence-measure een correlatiecoëfficiënt met de WER van -0.413.

Er is een foutje in de output van de herkenner ontdekt, waardoor de berekende confidencewaarden per zin erg hoog liggen. De *relatieve score* mag theoretisch niet boven 0 uitkomen, maar in de praktijk komt ongeveer 30% van de scores boven 0 uit. Hoewel experimenten hebben aangetoond dat de gekozen methode toepasbaar is voor monitoring, staat het vanwege de foutjes, niet onomstotelijk vast dat deze methode ook echt de beste van de twee is.

Aanbevelingen

In 4.3.3 is opgemerkt dat er er wellicht een andere confidence-methode kan worden ontwikkeld die nog beter is, door combinatie van de twee geteste methoden. Voor verder werk is dit een interessante opgave. Daarnaast is er een methode die uitgaat van gemiddelde duur van een foneem, en een “lattice”-methode (zie voor beiden 4.1.2) waarnaar verder onderzoek gedaan zou kunnen worden.

Er is in dit hoofdstuk aangegeven dat er waarschijnlijk een foutje zit in de MAX-likelihood-berekening van de herkenner. Dit zou opgelost moeten worden waarna de experimenten opnieuw gerund moeten worden. Daarnaast is de AVG-prior-likelihoods berekening niet helemaal goed. Ook dit zou verbeterd moeten worden, maar is gezien de interne structuur van *shout* helaas niet mogelijk.

5. Teletekstondertiteling

In dit hoofdstuk staat beschreven hoe de vele regels ondertiteling gebruikt kunnen worden als referentie voor de herkenning. Zoals al vermeld is, is de ondertiteling vaak redelijk gelijk aan wat er gezegd is. Dat dit niet altijd zo is staat uitgelegd in par. 5.2.1, Problemen ondertiteling. Verder staat het onderzoek beschreven naar het antwoord op de vraag in hoeverre de ondertitelingen bruikbaar zijn voor monitoring.

5.1. *Inleiding*

De meest bekende ondertiteling is “open ondertiteling”, de ondertiteling van niet-Nederlandstalige films en interviews. Voor doven en slechthorenden is er ook “gewone” Nederlandse ondertiteling, “gesloten ondertiteling” genoemd. Deze wordt uitgezonden via teletekst (Hendrickx 2002; Adnergje 2006). De ondertiteling van teletekst bestaat al jaren, en kan via pagina 888 opgevraagd worden. Voorheen bestond de ondertiteling uit korte simpele zinnen of samenvattingen. Vóór 2004 was de ondertiteling een samenvatting van het journaal, maar na een test in 2003(NVVS 2003) wordt er vanaf 2004 min of meer letterlijke ondertiteling uitgezonden. “*Hoewel de snelheid en de hoeveelheid van de tekst hiermee sterk verhoogd is, blijkt het goed te volgen”*(Adnergje 2006).

Ondertiteling in de literatuur

In de literatuur is niets gevonden over de toepassing van teletekstondertiteling (gesloten ondertiteling) als referentie van een herkenning, zoals in dit onderzoek. De gevonden toepassingen die gebruik maken van teletekst zijn wel gericht op de koppeling van spraakherkenning en ondertiteling, maar dan om de ondertiteling automatisch onder videosegmenten te plaatsen. De herkenning wordt vergeleken met de ondertiteling, en bij voldoende gelijkenis wordt de herkenning op die bepaalde plaats gezet (Huang, Hsu et al. 2003).

Ondertiteling in dit onderzoek

De ondertiteling die in dit onderzoek gebruikt wordt, is de gesloten ondertiteling. Deze is aangeleverd als teletekst-bestand in xml-formaat. Het bestand is gestructureerd per onderwerp van het journaal. Er is per regel ondertiteling, naast de tekst ook de begintijd en eindtijd gegeven (zie tabel 5.1).

```
<story start="20.00:00" end="20.02:56" id="0">
<p start="20.00:00" end="20.00:01">NOS, Journaal.</p>
<p start="20.00:04" end="20.00:08"></p>
<p start="20.00:09" end="20.00:13">De Grave wil inzet mariniers tegen
<p start="20.00:15" end="20.00:18">En Tilburg een paradijs voor saxof
<p start="20.00:22" end="20.00:23">Goedenavond.</p>
<p start="20.00:25" end="20.00:29">In een tv-toespraak heeft presider
<p start="20.00:32" end="20.00:36">Hij verbood 2 groepen die volgens
```

Tabel 5.1 Een praktijkvoorbeeld van een ondertitelingbestand in xml-formaat. Bovenin is aangegeven dat de “story” begint, plus de tijd dat het blok duurt (van 20.00u tot 20.03u). Een story betekent een journaal-onderwerp, in dit geval de opening en het eerste onderwerp. Daarna is per regel aangegeven van wanneer tot wanneer deze in beeld moet zijn, en de inhoud (tekst).

De ondertiteling kan gezien worden als een soort transcriptie. In de ideale situatie is de teletekstondertiteling letterlijk wat er gezegd is. Daarom zou het een uitstekende referentie moeten zijn. Het idee is dan ook om de prestatie van het systeem te meten zoals met transcriptie gebeurt. Er wordt dan een Word-Error-Rate (WER) bepaald voor alle segmenten, waardoor monitoring kan plaatsvinden voor het hele journaal.

Onderzoeksopzet

In dit onderzoek is gekeken naar de kwaliteit van de ondertiteling, en in hoeverre het bruikbaar is als goede referentie (REF). In de eerste plaats wordt gekeken naar de koppeling tussen herkenningen en ondertitels (die lopen niet altijd parallel). Daarna wordt geprobeerd met de ondertiteling-referentie (OT_REF) een betrouwbare WER_OT (Word-Error-Rate volgens de OT_REF) te bepalen die lijkt op de "echte" WER volgens transcriptie. Het is de bedoeling dat deze WER_OT (ook een percentage tussen 0% en 100%) dezelfde waarden of hetzelfde gedrag vertoont als de echte WER, waardoor deze bruikbaar is voor monitoring.

5.2. *Koppeling met hypothese*

Meestal komt de tekst in de ondertiteling behoorlijk overeen met hetgeen de journaallezer zegt. Ook zijn de teksten meestal in beeld op het moment dat iets wordt gezegd. Dit is echter niet altijd zo. Er zijn een aantal factoren die er voor zorgen dat de ondertiteling toch slechter is dan de handmatige transcriptie.

5.2.1. Problemen ondertiteling

De gesloten ondertiteling van journaals zijn sinds 2004 gebaseerd op de autocue van de nieuwslezer. De audio en ondertiteling waarmee de herkenner getraind is, en waarmee bijna alle onderzoeken op de UT worden gedaan zijn nog van voor 2003. In die tijd was de ondertiteling nog een samenvatting van wat de nieuwslezer had gezegd. Deze samenvatting is niet echt consequent. Soms is het een echte samenvatting van meerdere zinnen, meestal worden echter enkele belangrijke zinnen letterlijk gegeven, en de rest weggelaten.

Naast de "samenvatting" van de tekst van de nieuwslezer, bestaat de gesloten ondertiteling uit de open ondertiteling van niet-Nederlandstalige fragmenten (rapportages en niet-live interviews), en verder samenvattingen van Nederlandstalige fragmenten. Deze laatste wordt ter-plekke gemaakt en verzonden(Adnergje 2006). In alle verschillende onderdelen van de ondertiteling komen fouten voor. Deze zijn hieronder beschreven.

Nieuwslezer-tekst

Hoewel deze tekst grotendeels letterlijk is wat de nieuwslezer zou moeten zeggen, is dit vaak niet het geval. De nieuwslezer leest vaak niet perfect. Hij/zij kan zich verspreken, "eh" invoegen of woorden in een andere volgorde uitspreken ("dit moet men gedaan hebben" of "dit moet men hebben gedaan"). Deze oorzaken zorgen voor lagere kwaliteit van de referentie.

Het is duidelijk dat de delen van de "samenvattingen" die niet letterlijk zijn, of waar delen zijn weggelaten, geen goede referentie (OT_REF) opleveren.

Anderstalige audio

Deze audio zal geen bruikbare herkenning opleveren. De herkenning zal hooguit akoestisch lijken op wat er gezegd is, maar slecht lijken op de ondertiteling. Bij de experimenten in dit onderzoek zijn alle audiofragmenten met anderstalige audio verwijderd uit de bestanden. Bij toekomstige automatische verwerking zijn deze audio wél van toepassing. Monitoring kan helpen deze fragmenten op te sporen.

Interviews

Deze worden "live" gehouden. Hoewel het onderwerp en de vragen meestal bekend zijn, zullen de antwoorden van de geïnterviewde journalist veel audio in korte tijd brengen. Deze moet handmatig en snel omgezet worden naar teletekst. Dat gebeurt dan als samenvatting. Hierin staan vaak wel enkele belangrijke woorden in, maar als transcriptie is het slecht bruikbaar.

Formaat/hoeveelheid

Van het journaal van 12 januari '02 zijn 378 regels herkenning. In het xml-bestand met teletekst van die dag zijn er slechts 132 regels. Dit geeft al aan dat de toepasbaarheid van de ondertiteling niet groot is. Voor de monitoring zal de ondertiteling hooguit een rol kunnen spelen bij de segmenten waarvan ondertiteling beschikbaar is.

5.3. *Onderzoek*

Uit de vorige paragraaf is naar voren gekomen dat er de ondertiteling niet altijd overeenkomt met de transcriptie. Desalniettemin lijkt het vaak op de herkenning en zou het toch bruikbaar kunnen zijn in de gevallen dat er een ondertiteling beschikbaar is. In deze paragraaf wordt de koppeling van ondertiteling en herkenning besproken. Daarnaast worden scoringscoringsmethoden voorgesteld om de WER te bepalen.

5.3.1. Koppeling

De herkenning bestaat uit tekst en begin- en eindtijd. De ondertiteling heeft dat ook, en daarnaast nog informatie over de onderwerp-blokken waarin het zich bevindt. Om iets te kunnen zeggen over de prestatie van de herkenner volgens de ondertiteling, moet de herkenning gekoppeld worden aan de ondertiteling. De volgende problemen doen zich voor bij de koppeling. Hiervoor worden HYP en ondertiteling gezien als verzamelingen losse zinnen.

Problemen koppeling

- Alignment – Oftewel het linken van een zin uit de HYP met eentje uit de ondertiteling. Welke ondertitelingen hoort bij welke herkenning?
- Verschil in hoeveelheid – De ondertitelingen zijn soms per halve zin, soms meerdere zinnen. De herkenning is in het ideale geval per zin. Samenvoeging leidt tot grotere ondertiteling dan herkenning. Hoe wordt daarmee omgegaan?
- Scoring – De uitgesproken tekst is bijna nooit letterlijk hetzelfde als de ondertiteling. De originele scoring is daarom niet meer werkbaar omdat er teveel foutgerekend zal worden.
- Volgorde – In hoeverre wordt verwisseling van woorden binnen een zin goedgerekend?

Hieruit zijn de onderzoeksonderwerpen *alignment* en *scoringsmethode* ontstaan. Deze worden in de volgende hoofdstukken besproken.

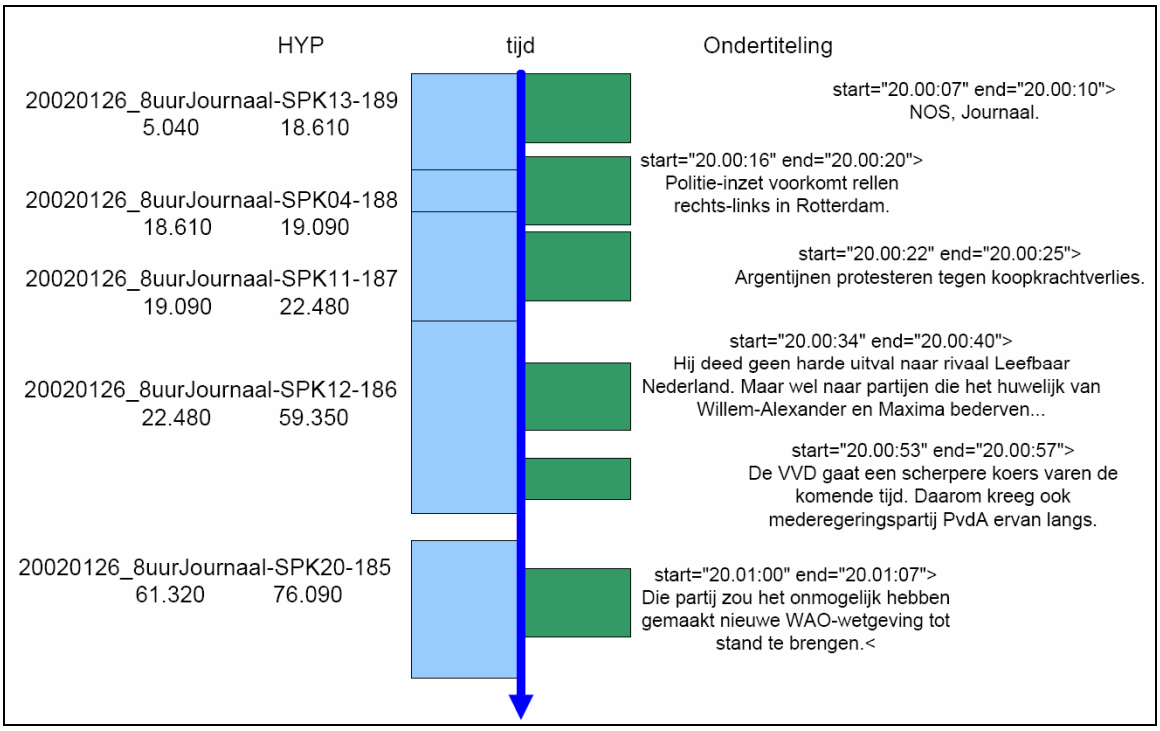
5.3.2. Alignment

Om de HYP te vergelijken met de ondertiteling (de OT_REF) is het nodig te weten welke ondertiteling bij welk stukje herkenning hoort. Daar deze nooit helemaal synchroon lopen, en er eerder genoemde problemen met de ondertiteling zijn, is er alignment nodig. Alignment is al beschreven in hst 2.2.2, "Alignment". Daar wordt met de term alignment, een methode bedoeld die twee zinnen correct koppelt (matching per woord). Bij teletekst-koppeling betekent alignment dat hele bestanden worden gekoppeld (matching per zin of segment). Het alignment-mechanisme koppelt de juiste HYP-zinnen met één of meerdere REF-zinnen (ondertitels).

De overwogen methoden voor alignment zijn op basis van tijd en oplijning. Aligment op basis van tijd wil zeggen dat er alle ondertitelingen die qua tijd overlappen met de herkenning, aan elkaar worden geplakt tot één referentiezin. Dit is vergelijkbaar met alle ondertitels die tijdens de spraak in beeld zijn geweest. Bij oplijning wordt niet gekeken naar de bekende tijdmarkeringen van HYP en REF_OT, maar worden de ondertitelingen in volgorde gekoppeld aan herkenningen. Dit koppelen houdt in dat elke zin uit de HYP met de best gelijkende zin uit de ondertiteling gematcht wordt, mits de volgorde van zinnen hetzelfde blijft.

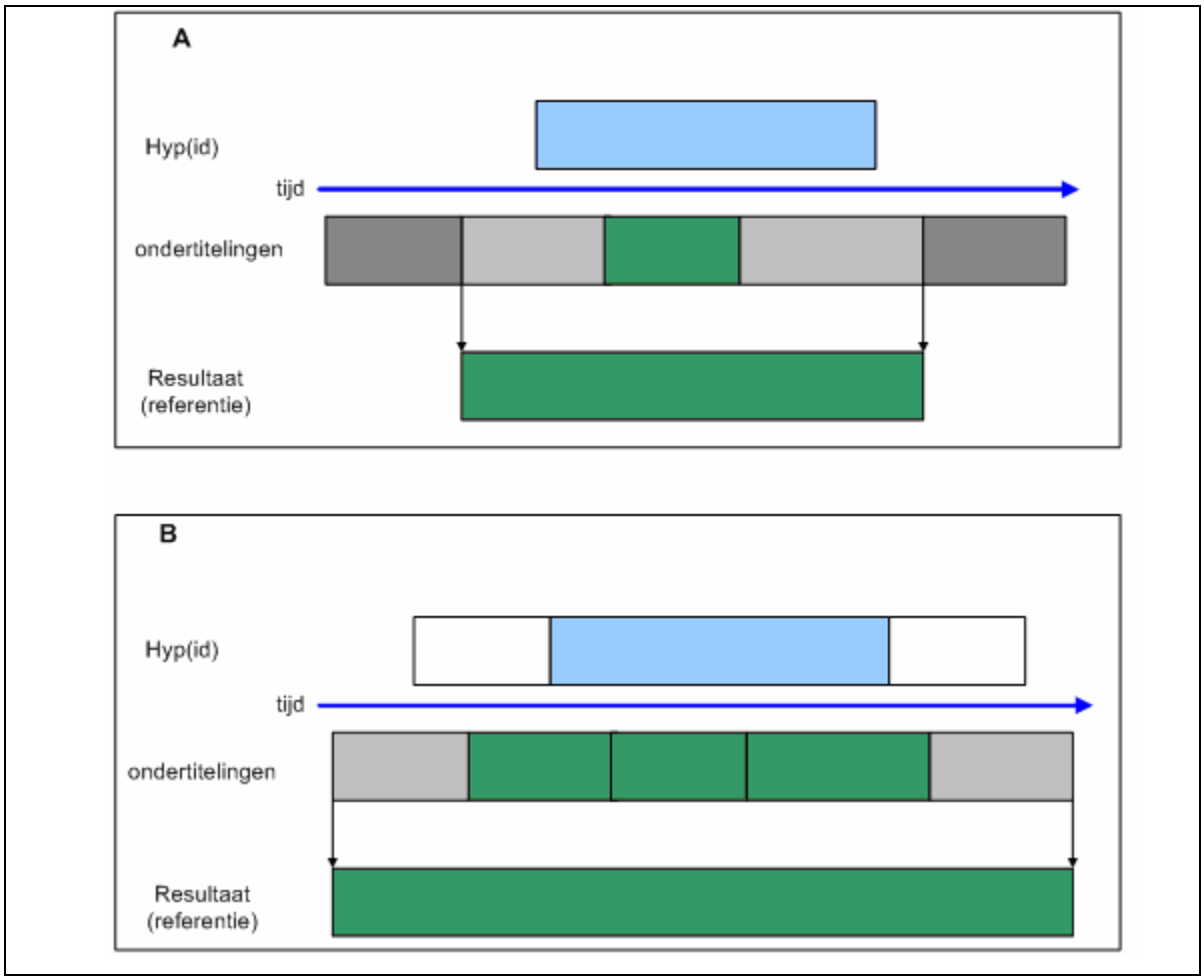
Alignment op tijd

Van zowel herkenning als ondertiteling is bekend van wanneer tot wanneer het duurt (zie tabel 5.2). Het is echter bijna nooit het geval dat de tijdmarkeringen van de HYP gelijk zijn aan die van één bepaalde ondertiteling.



Tabel 5.2 In dit figuur is te zien dat de tijden van de HYP-segmenten niet overeenkomen met zinnen uit de ondertiteling. Links vijf ID's met hun respectievelijke begin- en eindtijden, rechts de begin- en eindtijden van de ondertitelingen met de tekst van de ondertiteling.

Voor de toepassing is het idee om HYP en ondertiteling te vergelijken als ze ongeveer rond dezelfde tijd beginnen en eindigen. "Ongeveer dezelfde tijd" blijkt een moeilijk begrip. Voor het implementeren moeten er concrete voorwaarden aan gesteld worden. Na initieel onderzoek is gebleken dat het niet eenvoudig is om 1-op-1 herkenningen aan ondertitelingen te koppelen. Vaak vallen de ondertitelingen net buiten de begintijd en eindtijd van de herkenning. Daarom zijn de voorwaarden voor koppeling ruim genomen. Het nadeel van ruimere voorwaarden is dat veel onbruikbare zinnen aan de REF worden toegevoegd. Dit is een kwaliteit/kwantiteit-dilemma waarvoor een balans gevonden moet worden. In tabel 5.3 is te zien welke voorwaarden gekozen zijn ter onderzoek. Per HYP-segment worden de ondertitelingen bekeken. Alle ondertitelingen die in tijd overlappen met het HYP-segment, worden samengevoegd tot referentiezin met hetzelfde ID als het HYP-segment.



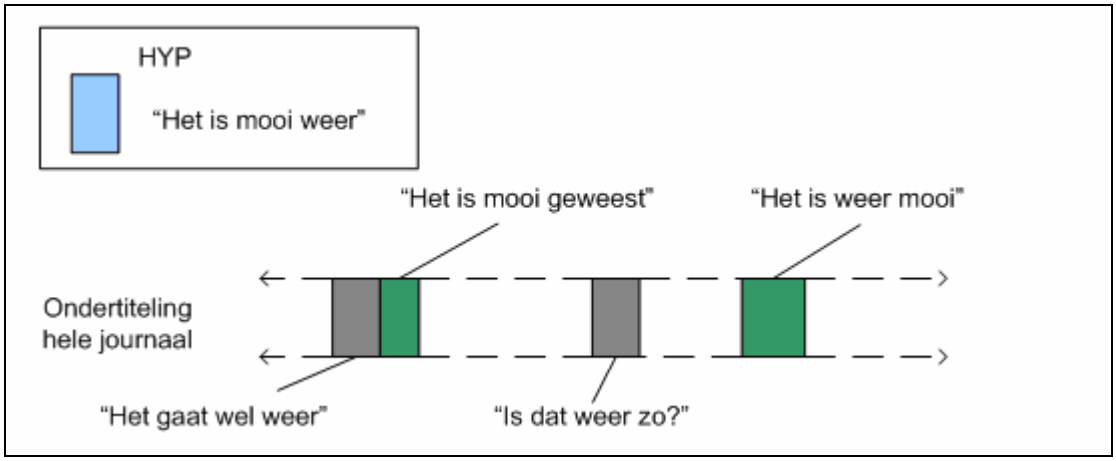
Tabel 5.3 De beide methoden om alignment op basis van tijd te doen. Bij methode A wordt de referentie samengesteld uit ondertitelingen die op minstens één moment overlappen met de tijden van het HYP-segment. Methode B laat ook nog de ondertitelingen toe tot de referentie als ze overlappen met het gebied van de marge van 2 seconden van het HYP-segment.

Bij de eerste onderzoeken (en handmatige controle) is gebleken dat er af en toe ondertitelingen zijn die niet overlappen met het bijbehorende HYP-segment. Hiervoor is koppel-methode B tabel 5.3 bedacht. Deze is nog ruimer dan de eerste, en heeft een marge van 2 seconden aan voor-en achterkant. Hierdoor kunnen ondertitels die nèt niet in beeld zijn geweest tijdens het uitspreken hiervan, toch als REF dienst doen. Het levert wel veel langere REFs op dan de eerste methode. Hierdoor werken de originele scoringsmethoden (met Sclite) niet meer. In de paragraaf 5.3.3, Scoring is hierover meer geschreven. Er is handmatig gekeken naar de resultaten om te controleren of er nog steeds ondertitels zijn die buiten de tijdmarkeringen vallen en daardoor niet gematcht worden aan hun spraak-segment. In dat geval zou er nog een ruimere categorie bedacht moeten worden. Er is niets gevonden. De experimenten worden dus gedaan met twee soorten referenties.

Alignment door oplijning binnen journaal

In tabel 5.4 is een indicatie voor de toepassing van oplijning te zien. De HYP-zin is de kandidaat, er wordt gezocht naar de beste ondertiteling om te matchen. De beste ondertiteling heeft de meeste overeenkomstige woorden, en liefst ook in dezelfde volgorde. De figuur geeft aan dat er vaak veel ondertitels zijn met dezelfde woorden. Dit komt waarschijnlijk omdat de ondertiteling van voor 2003 is, en grotendeels samenvatting is. Er wordt gezocht op de

overeenkomstige woorden, maar omdat maar een klein gedeelte van de HYP-zin voorkomt in de ondertiteling, zijn er meer ondertitels die redelijk lijken op de HYP-zin.



Tabel 5.4 Opzoeken van correcte ondertiteling door oplijning. De ondertitels die het meest lijken op het HYP-segment (linksboven) zijn kandidaten voor de correcte ondertiteling. De groene delen zijn kandidaten, de grijze iets minder en de rest van het journaal is geen kandidaat.

De oplijning zoals hierboven weergegeven geeft aan waar de problemen zitten. De herkende zin is "Het is mooi weer". Deze zin komt niet helemaal voor in de ondertiteling. Er zijn vier segmenten die gekoppeld zouden kunnen worden, maar twee zijn het meest waarschijnlijk. Het is echter lang niet zeker dat de ondertiteling een van deze is.

Het eerste probleem van de koppeling is dat de woorden "Het", "is", "mooi", "weer" verspreid kunnen voorkomen in het journaal. Samenvoeging van segmenten kan ertoe leiden dat er een grote referentiezin ontstaat met daarin de verschillende woorden. Een oplossing is om een tijd-beperking te gebruiken waarbij de ondertiteling ongeveer dezelfde lengte moet hebben als de HYP (bijvoorbeeld 0.5 tot 2x die lengte is in de praktijk *ongeveer*).

In (Huang, Hsu et al. 2003), al eerder genoemd, zijn verschillende oplijningalgoritmen besproken. Hierbij wordt echter uitgegaan van perfecte ondertiteling, zodat 1-op-1 koppeling mogelijk is, en meestal per HYP maar één ondertiteling kandidaat is om te matchen. Hier is geen sprake van perfecte ondertiteling. Daardoor wordt oplijning in dit systeem erg moeilijk. Er is een mogelijkheid om minder last te hebben van de "stopwoorden", de woorden die veel zinnen voorkomen en daarmee de oplijning moeilijker maken. Er is namelijk een door het TNO samengestelde lijst met stopwoorden. Als deze automatisch uit HYP en ondertiteling verwijderd worden, zouden alleen nog belangrijke woorden ("keywords") hoeven te vergelijken om tot een goede oplijning te komen.

Gekozen methode: tijd

Hoewel er uitgewerkte oplijningalgoritmen bekend zijn, is na kort onderzoek beslist dat de methode niet wordt toegepast voor dit onderzoek. Dit is omdat die oplijningalgoritmen uitgaan van een ondertiteling die erg goed is en letterlijk voorkomt in de herkenning. Dit is hier niet het geval, waardoor het oplijningalgoritme onbetrouwbaar wordt. Omdat de ondertiteling vaak kleiner in formaat is dan de HYP, zouden ondertitelingen samengevoegd moeten worden. Dit is niet toepasbaar binnen de bestaande algoritmen. De methode waarbij stopwoorden worden weggelaten is interessant. Er is gekeken naar het verwijderen van stopwoorden, maar wegens tijdgebrek is het implementeren van een oplijningalgoritme niet verder onderzocht.

Voor de toepassing binnen dit afstudeeronderzoek is de methode toegepast die het makkelijkst te implementeren was: alignment op tijd. Hiermee wordt aan elke herkenningszin oftewel HYP, de beste referentiezin (REF) gekoppeld. Door samenvoeging van meerdere ondertitels (met de de

marge 2 seconden-methode) is de REF-zin door langer geworden dan de HYP-zin. Bij het scoren moet hiermee rekening gehouden worden.

5.3.3. Scoring

Zoals uitgelegd in het hoofdstuk “Spraakherkenning” wordt de Word-Error-Rate (WER) bepaald door een hypothesezin met een referentiezin te vergelijken. Dit gebeurt nu meestal met de oplijning-tool Sclite. Het Sclite-programma past een oplijningalgoritme toe dat twee zinnen die op elkaar lijken goed kan vergelijken en direct de scores uitrekent (zie paragraaf 2.2.2,Alignment).

De gekozen align-methode “op tijd” kan grote REFs opleveren. Een andere methode (met nog grotere REFs) is om het hele journaal als REF te nemen. Sclite heeft echter moeite met zinnen van verschillende grootte. Alle woorden die niet ge-aligned kunnen worden, worden foutgerekend (zie pagina 46, ”Probleem Sclite”). Hierdoor is het niet mogelijk té grote referentiezinnen toe te laten, en een verdeling vooraf (de globale aligning) noodzakelijk is.

```
id: (20020112s003-012)
Scores: (#C #S #D #I) 3 1 5 0
REF:  DE gevangenen mogen HUN GELOOF  uitoefenen EN WORDEN GELUCHT
HYP:  ** gevangenen mogen *** GELOVEN uitoefenen ** ****** *******
Eval: D                   D   S                   D  D      D
```

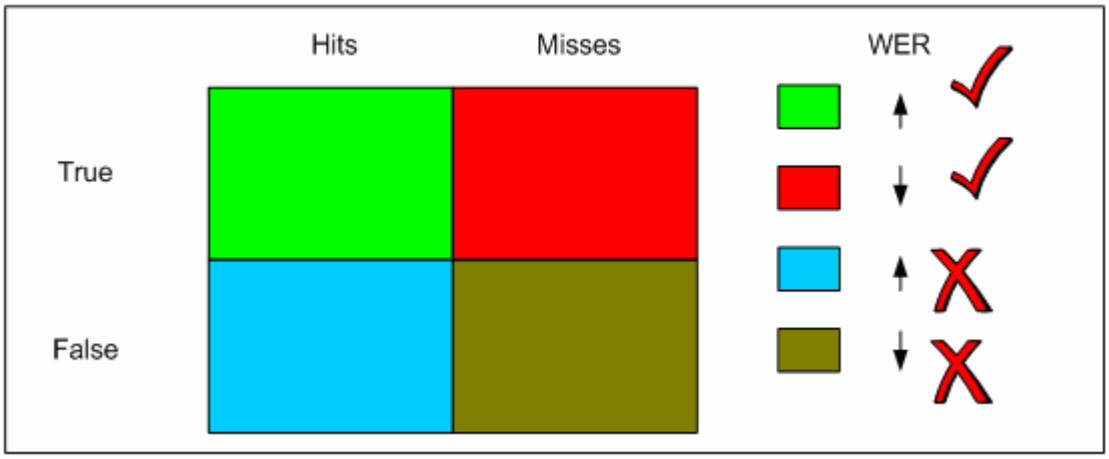
Tabel 5.5 Sclite-output. De samengevoegde ondertiteling staat naast “REF:”, de herkenning daaronder (naast “HYP:”). Aangezien de ondertiteling langer is dan de herkenning, oordeelt Sclite dat de woorden niet herkend zijn, en rekent die woorden fout (“D” van deletion in de onderste rij).

Bij gebruik van ondertiteling is de lengte van de referentie onbekend, en zeker niet gelijk aan die van de hypothese. Meestal is de referentie langer, wat leidt tot veel deletions aan begin en eind van de zinnen. In tabel 5.5 is hiervan een bescheiden voorbeeld gegeven, meestal is de REF veel groter. In de inleiding is al genoemd dat hiernaast ook nog verwisselingen plaatsvinden of woorden simpelweg missen uit de referentie. Voor al deze problemen is de normale Sclite-toepassing ongeschikt omdat er dan teveel woorden onterecht foutgerekend zullen worden en de WER van die herkenningen daardoor te hoog zal zijn. In het voorbeeld is 6 vd 9 woorden fout, oftewel een WER van 67%.

Vanwege dit soort problemen is een nieuwe scoringsmethode nodig die rekening houdt met de structuur van de nieuwe referentie. Er is gezocht naar methoden die een WER opleveren die realistischer is, waarbij deze controle zowel met de hand is, als statistisch met SPSS.

Herkenningsfouten algemeen

Bij het scoren van de HYP, zonder transcriptie, is het onmogelijk om een helemaal correcte WER te bepalen. De methoden die hieronder beschreven worden, hebben allemaal ofwel tot gevolg dat er onterecht volgordeverwisselingen foutgerekend worden (*sclite-output bewerken*) danwel dat onterecht deletions goedgerekend worden(*Frequentie tellen*). Dit probleem kan beschouwd worden als een zoekprobleem (zoals bij internet-zoekmachines). Het gaat hierbij om “hits” en “misses”. Het is de bedoeling dat de hits (goedgerekend) en misses (foutgerekend) zoveel mogelijk terecht zijn.

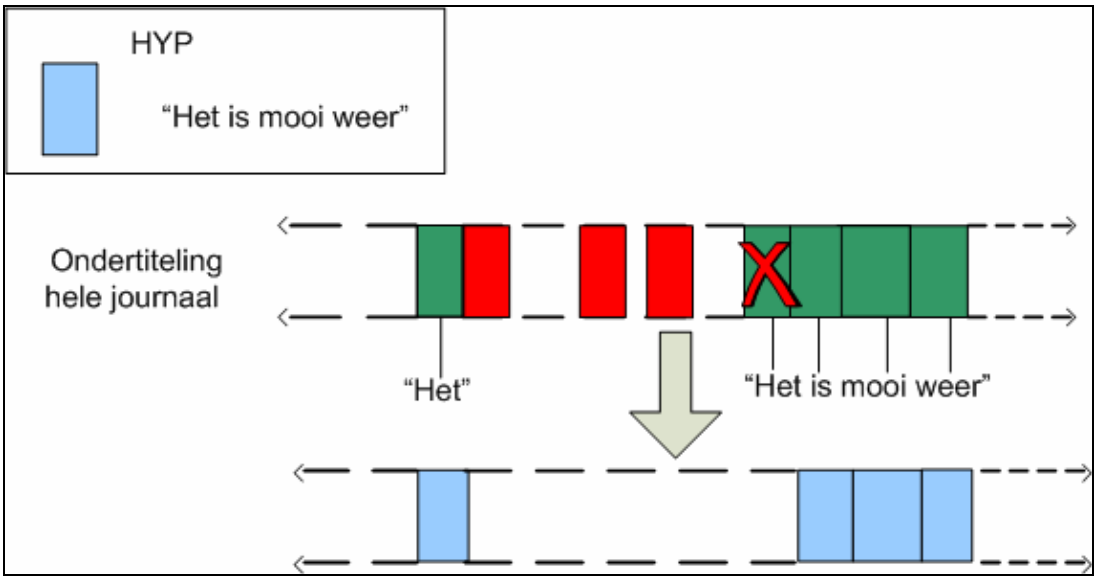


Tabel 5.6 Hits&misses. De vier categorieën zijn links aangegeven, rechts per categorie het gevolg voor de WER, met een vinkje als het gewenst is, en een kruis als het ongewenst is. De false hits en false misses zijn onzekerheden voor correct bepalen van de WER (voor monitoring dus) en moeten zo laag mogelijk gehouden worden.

Bij dit onderzoek wordt geprobeerd de methode te vinden die de minste **false** hits&misses tot gevolg heeft. Het is echter niet mogelijk om dit per woord te doen, er wordt slechts gekeken naar de WER van de zin. De WER van de methode wordt vergeleken met de “echte” WER van de transcriptie. Dit heeft tot gevolg dat niet meer apart naar de true&false (terecht of onterecht) gekeken wordt, maar alleen nog naar de hits&misses, die zijn bepalend voor de WER. De methode die globaal evenveel false hits als false misses negeert, en dus een WER bepaalt die goed zal correleren met de “echte” WER, is de beste methode. Het is zelfs nog goed als de methode een te hoge of te lage WER bepaalt, als het maar samenhangt met de “echte” WER. Een methode die dus alleen false hits of alleen false misses “aanpakt”, kan nog steeds goed werken.

Probleem Sclite

De oplijningmethode van Sclite (zie 2.2.3) is zeer geavanceerd, maar kijkt niet naar de afstand tussen woorden. Als de referentie groot is, probeert Sclite de HYP-woorden in de goede volgorde te matchen aan woorden uit de referentie, zie tabel 5.7.



Tabel 5.7 Sclite op kleine hyp en ref. Nadat “Het” is gevonden, wordt verder gezocht naar “is mooi weer”. De blauwe blokken onderin geven de uiteindelijke mappng weer.

De door het systeem gekozen “oplijning” is degene met de laagste WER. Bij een grote REF zoals in het figuur waarin de juiste woorden voorkomen, worden alle andere woorden foutgerekend. De ene oplijning heeft “Het” helemaal links (zoals in de figuur), de andere heeft “Het” gekoppeld aan de tweede “Het” in de REF. De beide oplijningen hebben dan evenveel foute woorden volgens Sclite. In dit geval kan een oplijning gekozen worden die niet de beste is.

Sclite-output bewerken

In tabel 5.5 is te zien dat de deletions een scheef beeld geven. Gebruik van de ondertiteling als referentie (met overlap, zie “Alignment op tijd”, pagina 42) zorgt in bijna alle gevallen voor extra deletions aan begin en eind van de referentie. Hiermee is de originele Sclite-scoringsmethode niet meer betrouwbaar. Het aantal false-misses zou veel te groot worden om een correcte WER te bepalen.

Een oplossing hiervoor zou zijn om alle deletions niet mee te tellen. Deze methode is te rigoreus want er zijn ook juiste deletions. Met methoden als deze wordt juist de kracht van Sclite aangetast. Een oplossing zou zijn om alleen de insertions aan de buitenkant (links en rechts van de gemapte REF, zonder *** dus) niet mee te tellen.

Dan zijn er nog gevallen waarbij er weinig ondertiteling aanwezig was, waarbij er meer HYP is dan REF, een situatie die inde praktijk weinig voorkomt. In dit geval levert Sclite veel insertions. Deze zouden op dezelfde manier weggehaald moeten worden.

Het kan voorkomen dat er juist een correcte deletion of insertion aan de rand van zo’n cluster staat. Dit is echter zeldzaam vergeleken met het aantal keren dat dit niet zo is, vandaar dat deze methode toch gerechtvaardigd is.

Frequentie tellen

Hier wordt niet meer gekeken naar de volgorde van de woorden, maar alleen naar de frequentie dat een woord in zowel de HYP-zin als de REF-zin voorkomt.

REF: DE gevangenen mogen HUN GELOOF uitoefenen EN WORDEN GELUCHT
HYP: ** gevangenen mogen *** GELOVEN uitoefenen ** ****** *******

HYP	HYP#	REF#
gevangenen	1	1
mogen	1	1
geloven	1	0
uitoefenen	1	1

Score WER = 1 - 3/4 = 1/4

Tabel 5.8 dezelfde HYP en REF als voorbeeld tabel 5.5, maar nu met de methode “frequentie tellen”. In het midden is te zien hoe vaak de woorden voorkomen in HYP en REF, waarna onderaan te zien is dat 1 op de 4 woorden foutgerekend wordt.

Bij frequentietellen wordt zeer “coulant” gescoord. De WER ligt altijd lager dan de methode volgens de sclite-output. Het idee is dat van alle verschillende woorden die in de HYP voorkomen, de frequentie in zowel HYP als REF wordt bijgehouden. Naar de andere woorden wordt niet gekeken. Deze WER is de som van frequenties in de REF, gedeeld door de som van de frequenties in de HYP.

Deze methode moet vooral een oplossing zijn voor de volgordeverwisseling en slechte oplijning. Daarnaast worden hiermee ook de deletions aan de randen mee weggewerkt. Nadelen zijn dat “Echte” deletions ook worden weggewekt, wat onterecht scoreverbetering tot gevolg heeft. Bij een grote referentie kan het ook voorkomen dat een woord op een totaal verkeerde

plaats in de referentie voorkomt, en dus onterecht goedgerekend wordt (zie tabel 5.9). Daarnaast kan een hele andere zin met toevallig veel dezelfde woorden hoog scoren. Omdat simpele woorden als "*de, als, beide, bijna, boven"* in veel zinnen voorkomen levert deze methode veel onterecht goedgerekende woorden.

```
id: (20020112s001-001)
Scores: (#C #S #D #I) 4 2 1 1
REF:  geen plaats VOOR MOSLIMS die VERANTWOORDELIJK ******* zijn
HYP:  geen plaats VAN  ******* die VOOR             WOORDEN zijn
Eval:             S    D           S                I
```

Tabel 5.9 In dit voorbeeld is "voor woorden" herkend, terwijl "verantwoordelijk" gezegd was. Omdat "voor" op een andere plaats in de REF voorkomt, wordt dit woord volgens het frequentietellen toch goedgerekend.

Er is een stopwoordenlijst beschikbaar, al eerder genoemd, waarmee het mogelijk is om deze woorden te verwijderen uit zowel HYP als REF. Daardoor blijven de zogenaamde *keywords*, oftewel de belangrijke woorden over. Hiermee kan het aantal false hits geminimaliseerd worden.

De oplijningmethode en de frequentiemethode kunnen beide aangepast worden voor keywords. Na initiële experimenten met weinig resultaat is besloten deze methode niet verder uit te werken.

5.3.4. Uit te voeren experimenten

In dit hoofdstuk zijn de verschillende problemen besproken die komen kijken als ondertiteling gebruikt moet worden voor monitoring. Het eerste is de koppeling van herkenning en ondertiteling. Er is besloten door experimenten te bepalen welke methode met alignment op tijd het beste werkt: overlapping of overlapping met marge (zie: paragraaf 5.3.2, Alignment,Gekozen methode: tijd). Daarnaast worden experimenten gedaan om de beste scoringsmethode te bepalen: aangepaste Sclite of frequentietellen. Deze twee soorten experimenten zijn helaas niet los te doen, als er geen koppeling is, kan er ook niets gescoord worden. Daarom zijn er vier experimenten gedaan, beide koppelingsmethoden gecombineerd met beide scoringsmethoden.

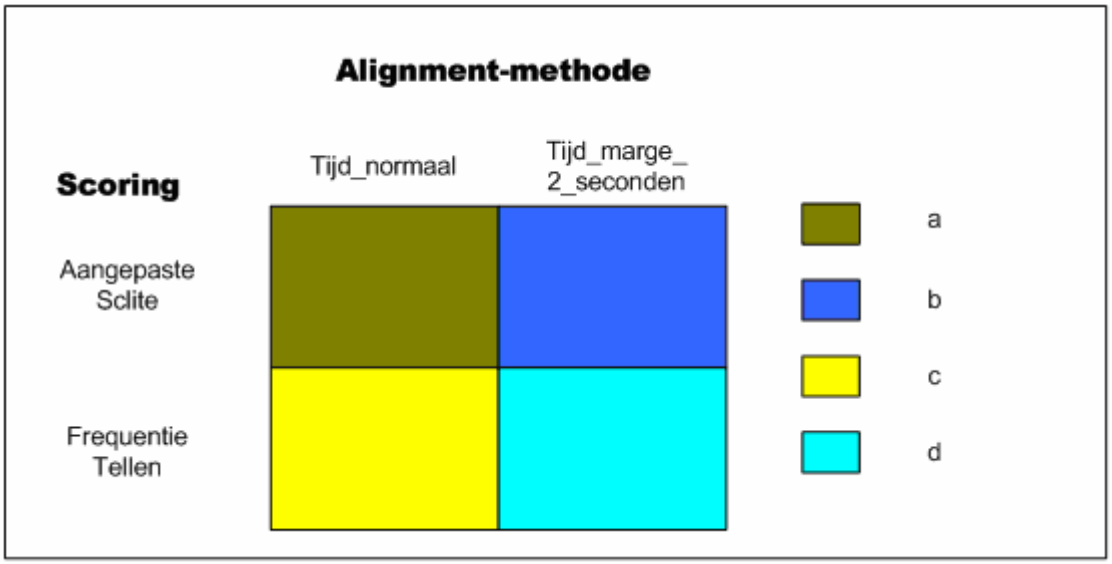
In bijlage A zijn alle scripts beschreven die gebruikt zijn voor de verschillende methoden.

5.4. *Uitvoer experimenten*

In deze paragraaf wordt verslag gedaan van de experimenten die zijn gedaan om alignment en scoringsmethode op elkaar af te stemmen. Hiermee is de methode bepaald die het best herkenningen van journaals kan monitoren.

5.4.1. Inleiding

Er zijn dus twee alignment-methoden en twee scoringsmethoden die onderzocht moeten worden. Uiteindelijk levert dat vier categorieën op, a t/m d. De categorie met de hoogste correlatie met de werkelijke WER is de winnaar en geeft antwoord op de onderzoeksvraag "hoe moeten HYP en ondertiteling als REF gekoppeld worden".



Tabel 5.10 De categorieën a t/m d staan voor de vier combinaties van methoden waarmee een WER_OT bepaald wordt. Boven de twee alignment-methoden, links de scoringsmethoden.

5.4.2. Aanpak

Zoals eerder genoemd zijn de experimenten gedaan op het journaal van 12 jan '02. Hiervan zijn de herkenning en transcriptie beschikbaar.

Door middel van de voorgestelde koppelingsmethoden zijn de OT_REF_normaal en OT_REF_marge_2_seconden bepaald. Dit is gedaan op de manieren zoals in tabel 5.3 is aangegeven. De categorie met marge van 2 seconden heeft een iets groter referentiebestand opgeleverd. Dit komt doordat er door de marge meer overlap is met de tijden uit de herkenningen. Om de methoden correct te vergelijken is gezorgd dat in beide HYP-bestanden de ID's dezelfden moeten zijn. Voor bepaling van de WER_OT zijn ook in de OT_REF-bestanden de overige ID's weggelaten. Ditzelfde is gebeurd in de originele REF (van de handmatige transcriptie), zodat de "echte" WER ook bepaald kon worden.

Dit heeft uiteindelijk 116 herkenningen ID's opgeleverd waarvan zowel herkenning, transcriptie, "normale" referentie op tijd, en referentie op tijd met marge van 2 seconden zijn.

	id	OT_norm1	OT_norm2sec	FreqT_n1	FreqT_n2s	hrReal
1	20020112s000-000	0%	0%	0%	0%	0%
2	20020112s000-002	38%	38%	38%	38%	38%
3	20020112s000-005	100%	100%	100%	100%	100%
4	20020112s001-001	22%	65%	22%	22%	0%
5	20020112s001-002	33%	33%	33%	33%	33%
6	20020112s001-005	71%	71%	57%	57%	36%
7	20020112s001-006	75%	75%	75%	75%	0%
8	20020112s001-013	100%	100%	100%	100%	46%

Tabel 5.11 Word-error-rates (WERs). De meest rechtse kolom geeft de echte WER aan. De andere kolommen zijn categorieën a t/m d van par. 5.4.1. Dit zijn de eerste acht van de 116 id's.

De originele transcriptie is de beste REF. Er wordt van uit gegaan dat de sclite-methode (zie 2.2.2, Alignment) de beste methode is om de herkenning hiermee te vergelijken. De WERs per segment, die hiermee bepaald zijn, zijn de norm, daarmee worden de nieuwe scores vergeleken. In de experimenten is voor de categorieën a t/m d, per ID de WER_OT bepaald. Deze is met SPSS

vergeleken met de “echte” WER. De methode met de hoogste correlatiecoëfficiënt over alle ID’s is de beste alternatieve methode.

5.4.3. Resultaten

In tabel 5.12 wordt een eerste analyse op de WERs gedaan. In de figuur zijn de gemiddelde WER, en de standaardafwijking te zien.

Report

Conf (Banded)		OT_norm1	OT_norm2sec	FreqT_n1	FreqT_n2s	hrReal
Total	Mean	58.17%	57.79%	51.84%	49.75%	36.22%
	N	116	116	116	116	116
	Std. Deviation	32.757%	31.993%	32.634%	32.178%	30.131%

Tabel 5.12 Een van de vele analytische reports binnen SPSS. Hier zijn per categorie a t/m d plus de “echte” WER, aangegeven wat de gemiddelde WER is. Verder is aangegeven dat de standaardafwijking iets boven de 30% ligt.

Bij tabel 5.12 moet gezegd worden dat de gemiddelde WER_OT niet gelijk is aan de totale WER_OT van het hele journaal is, maar de gemiddelde WER_OT per segment. De segmenten variëren in aantal woorden van 1 tot 26 woorden. Een zin met 11 woorden die helemaal goed is, leidt dus samen met een zin van 1 woord die helemaal fout is tot een gemiddelde WER van 50%. De daadwerkelijke WER_OTs zijn te zien in tabel 5.13.

WER_ot_a	WER_ot_b	WER_ot_c	WER_ot_d	WER_echt
57.52%	57.14%	48.60%	46.42%	32.76%

Tabel 5.13 De WER_OTs van de verschillende categorieën a t/m d + de WER van categorie “echt”.

Wat betreft de standaarddeviatie in tabel 5.12, deze ligt iets boven de 30%, wat inhoudt dat 68% van de 116 segmenten tussen de “gemiddelde” + 30% en “gemiddelde”-30% ligt. Dit geeft alleen aan dat de WERs nogal variëren. Het blijkt (overigens ook in tabel 5.12 te zien) dat de ondertiteling-REFs voor een nogal hoge gemiddelde WER zorgen. De REF van categorie “d” lijkt qua gemiddelde het meest in de buurt van de echte REF te komen.

Voor het bepalen van de correlatiecoëfficiënt wordt niet gewerkt met het “werkelijke gemiddelde”, maar met de WER per segment, dus onafhankelijk van het aantal woorden in een segment. In tabel 5.14 zijn de verschillende correlatiecoëfficiënten te zien.

Correlations

		OT_norm1	norm2sec	FreqT_n1	FreqT_n2s	hrReal
hrReal	Pearson Correlation	.584	.586	.610	.620	1
	Sig. (2-tailed)	.000	.000	.000	.000	
	N	116	116	116	116	116

Tabel 5.14 de correlatiecoëfficiënten van de WER_OTs van verschillende categorieën a t/m d met de “echte” WER. Het verschil is niet echt groot, maar categorie d lijkt het beste te correleren met de echte WER.

Het lijkt er op dat de categorie d de beste REF is, daar het met een correlatiecoëfficiënt van 0.620 de hoogste score heeft behaald.

5.5. *Conclusies*

De onderzoeken in dit hoofdstuk zijn gedaan om te bepalen in hoeverre ondertiteling bruikbaar is als alternatief voor de transcriptie, om een goede WER_OT te bepalen die hetzelfde gedrag heeft als de echte WER. De koppelmethode tussen HYP-bestand en ondertiteling-bestand heeft gezorgd voor een ondertiteling-REF die meestal meer woorden bevatten dan de HYP-zinnen. Hierdoor is het minder goed mogelijk gebleken om de (gebruikelijke) scorebepaling met Sclite te doen.

De methoden “frequentietellen” zal vaak woorden onterecht goed rekenen, maar het blijkt dat dit niet in extreem lage WER_OT’s resulteert. In tabel 5.11 is zelfs te zien dat gemiddeld gezien de WER toch lager ligt dan de WER_OT van deze methode. Dit komt waarschijnlijk omdat de ondertiteling veel slechter is dan de werkelijke transcriptie.

Uit de tabel 5.14 is af te leiden dat de beste methode het frequentietellen, in combinatie met de koppelingsmethode met marge van 2 seconden is. De correlatiecoëfficiënt met waarde 0.620 verschilt niet zo veel met die van de andere drie combinaties, maar het getal zélf is vrij hoog. Ook lijkt de gemiddelde WER_OT hierbij meest op de “echte” WER. Hiermee is aangetoond dat deze categorie zowel qua gedrag als qua gemiddelde waarde het meest lijkt op de WER. Er kan geconcludeerd worden dat deze combinatie van alignment en scorebepalen goed gebruikt kan worden voor monitoring. Er moet wel gezegd worden dat de ondertiteling niet voor alle segmenten beschikbaar is, dus daardoor niet het hele nieuws gemonitord kan worden.

5.5.1. Aanbevelingen

Als de prestatie van de herkenner beter wordt, of de ondertiteling beter overeenkomt met wat er gezegd is, moet er meer gedaan worden met alignment door oplijning. De herkenningen moeten eerst allemaal op de juiste plaats gezet worden in het journaal, voordat er scoring kan plaatsvinden. De ondertiteling van de nieuwe journaals (na 2003) zou al veel beter moeten zijn, waardoor de gemiddelde WER_OT lager zal liggen, en de dekking beter is, dus de correlatiecoëfficiënt nog hoger ligt. Wellicht dat met de betere ondertiteling de oplijningvariant voor koppeling van segmenten aan ondertiteling al mogelijk is. Met de huidige methode is de REF soms veel te groot, daardoor is deze onbetrouwbaar geworden. Mapping zal dan de oplossing zijn. Er zal ook gezocht moeten worden naar een Sclite-variant die afstand tussen de woorden minimaliseert. Toepassing hiervan zal wellicht nu al interessante toepassing binnen het monitoren mogelijk maken.

6. Training met behulp van monitoring

In dit hoofdstuk wordt het onderzoek naar hertraining en adaptatie met imperfecte data besproken. Er wordt uitgelegd hoe hertraining en adaptatie werken en welke experimenten zijn gedaan om de beste manier van *unsupervised adaptatie* te vinden.

6.1. *Inleiding*

De verschillende experimenten en onderzoeken die in deze scriptie zijn beschreven, hebben mogelijkheden aangereikt voor toepassing van monitoring. Voor training is audio nodig waarvan ook transcriptie is. Het idee is om een selectie van de herkenningen, de béste herkenningen, te beschouwen als transcriptie. Hoewel daar nog wel fouten in zitten is de hoop, dat het trainen van modellen op deze data betere herkenningen tot gevolg heeft.

Het doel is te onderzoeken welke toepassing het best te realiseren is: training of adaptatie. Uit het literatuuronderzoek is gebleken dat dit adaptatie is. De experimenten zijn gericht om iets te kunnen zeggen over de beste verhouding hoeveelheid en kwaliteit van de geselecteerde trainingsdata. Voor de experimenten worden enkele categorieën van verschillende kwaliteit en verschillende grootte gebruikt. Het is niet direct de bedoeling om grenzen af te bakenen, maar meer om een inzicht te krijgen in de mogelijkheden van adaptatie door gebruik van monitoring.

Daarnaast kan een succesvol adaptatieresultaat aantonen dat de methoden voor monitoring goed werken. Als blijkt dat het unsupervised (dus zonder dat een mens controleert) selecteren van de beste data goed werkt, zegt dat dat de interne confidence en/of de teletekst-methode de kwaliteit van de herkenningen goed weergeven.

6.2. *Onderzoek*

Training of hertraining houdt in dat het akoestisch model opnieuw wordt gemaakt op basis van de audio en bijbehorende tekst (zie paragraaf 2.1.3). Adaptatie betekent dat het model alleen wordt aangepast op basis van de nieuwe input. In dit hoofdstuk is de werking van beide methoden besproken. Daarnaast is uitgelegd dat unsupervised adaptatie de beste kans van slagen heeft, en hoe de experimenten voor de beste adaptatie zijn opgezet.

6.2.1. Literatuur over training en adaptatie

Voor dicteersystemen, is het noodzakelijk training of adaptatie te gebruiken. Deze systemen worden voor één persoon geoptimaliseerd. Zulke systemen moeten een lage WER garanderen. Bij dicteersystemen is het gebruikelijk dat de gebruiker eerst een aantal woorden en zinnen opleest van het scherm, zodat de computer de karakteristieken van de stem van de gebruiker kan bepalen. In (Thelen 1996) wordt geconcludeerd dat adaptatie de voorkeur heeft boven training. Training op beperkte data (enkele uren) heeft nog steeds tot gevolg dat sommige foneem-overgangen niet zijn uitgesproken, waardoor herkenning daarvan slecht is. Het beste werkt lange-termijn adaptatie waarbij enkele uren spraak gebruikt worden voor adaptatie. Dit gebeurt supervised (met de correcte referentie) en unsupervised (met de herkenning als enig controlemiddel), waarbij de categorie supervised duidelijk beter presteert.

(Burnett and Fanty 1996) bespreken experimenten over zeer korte-termijn-adaptatie. Het gaat hier om een supervised systeem waarbij het model geadopteerd wordt op de spraak van kinderen. De meeste akoestische modellen zijn getraind op volwassenen-spraak, vandaar dat de herkenning van kinderspraak meestal slecht is. De kinderen moeten één of enkele getallen uitspreken, waarna de herkenner op basis van deze beperkte input, het akoestisch model adapteert. Hier gaat het dus om korte-termijn adaptatie. Bij een klein lexicon blijkt dit goed te werken.

In (Giuliani and Brugnara 2006) wordt een interessante methode getoond waarin verschillende herkenners hypothesen geven. De woorden die in beide HYPs voorkomen worden als “juist” getypeerd, en daarmee wordt geadopteerd. Er wordt multiple-step adoptatie toegepast waarbij de output van de ene stap, de input van de andere stap is. Na deze eerste pass verbetert de prestatie het meest. Dit systeem is bruikbaar voor lopende spraak. Het is tevens vergelijkbaar met dit onderzoek. De ondertiteling zou dan gezien kunnen als de tweede herkenner.

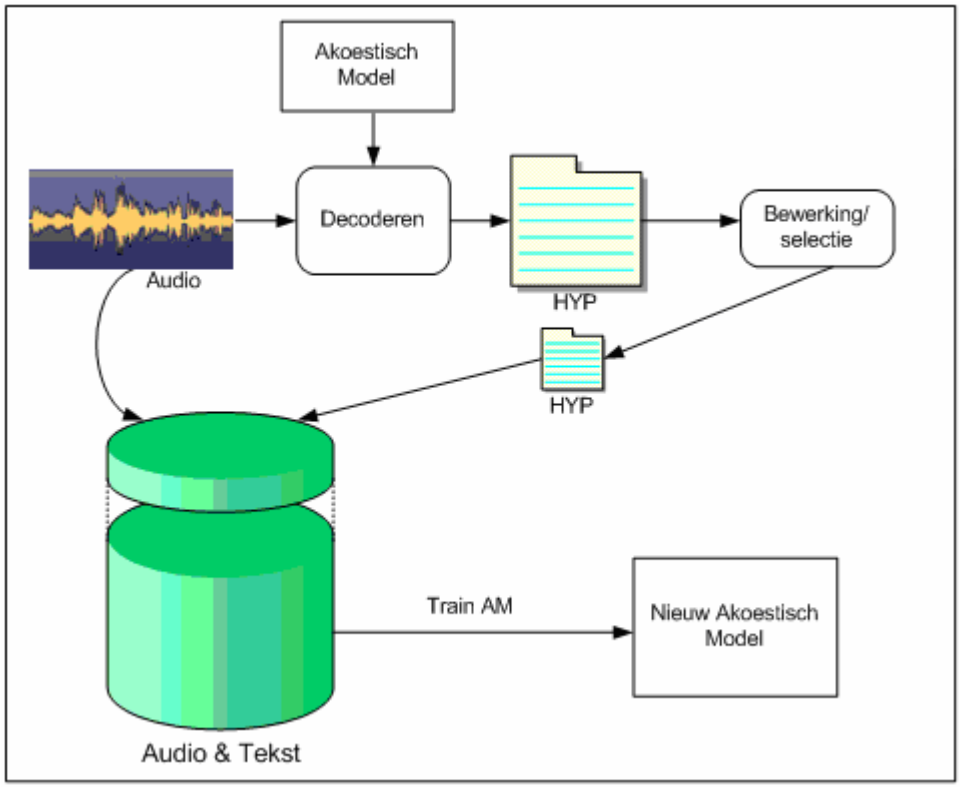
De eerste bronnen gaan over *speaker-adaptation*. Hierbij wordt gericht geadopteerd voor één persoon. Binnen een journaal spreekt veelal één persoon, maar het heeft ook andere spraak. Supervised adoptatie (met gebruik van transcriptie als referentie) wordt niet onderzocht in dit onderzoek. Het presteert wellicht beter, maar dit onderzoek richt zich vooral op de mogelijkheden van monitoring, en daaruit voortkomend unsupervised adaptation.

6.2.2. Training versus adaptatie

Van beide methoden wordt uitgelegd hoe het werkt. Daarna wordt besproken wat de meest bruikbare methode is voor het systeem.

Training

Zoals in 2.1.3, Training is uitgelegd, wordt de (her)training van het AM op een grote hoeveelheid trainingsdata gedaan. Als daar nieuwe data aan wordt toegevoegd, zoals het doel van dit onderzoek is, wordt het hele model opnieuw getraind op alle data, dus ook de oude. Alleen op de níeuwe data trainen is mogelijk, maar onverstandig, daar de herkenner met het nieuwe model dan alleen goed zal presteren op dat ene journaal. Dit is niet wenselijk voor de lange termijn. Het hertrainen van een akoestisch model is schematisch te zien in tabel 6.1.



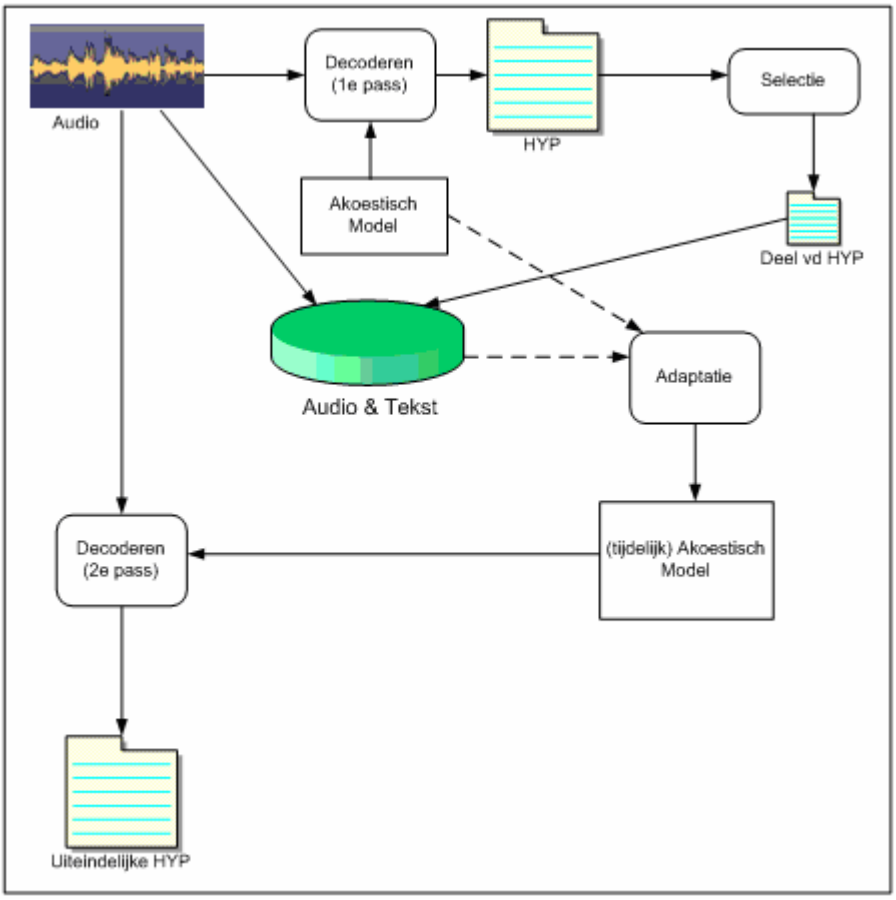
Tabel 6.1 De schematische weergave van unsupervised training. Waar oorspronkelijk door koppeling van audio en transcriptie een nieuw model getraind kon worden, wordt dat nu gedaan door een deel van de audio en een deel van de herkenning (HYP). De bewerking/selectie-unit selecteert de beste herkenningen.

Dit “Nieuw Akoestisch Model” kan weer gebruikt worden door de herkenner om nieuwe (en oude) audio te herkennen. In het plaatje van tabel 6.1 komt dit model dan op de plaats van het “oude” akoestisch model (bovenin). De *decoder* zal dan dit model gebruiken om de audio om te zetten naar een nieuwe HYP.

Adaptatie

Naast training is er nog adaptatie, wat lijkt op training. Adaptatie wil zeggen dat het akoestisch model **aangepast** wordt. De parameters van het akoestisch model worden als het ware verschoven in de richting van de *feature vectors* van de data die ter adaptatie aangeboden wordt. Adaptatie heeft als voordeel dat het niet uitmaakt op hoeveel data het akoestisch model getraind is. In tegenstelling tot training is het resultaat direct te zien.

De unsupervised adaptatie die hier beoogd wordt, houdt in dat na de eerste herkenning van een journaal, het akoestisch model automatisch wordt aangepast, en nogmaals een herkenning wordt gedaan. Adaptatie vindt dus plaats binnen één herkenning (van een heel journaal). Dit is effectief als het journaal bepaalde akoestische kenmerken heeft die anders zijn dan de gemiddelde journalen. Het akoestisch model kan dan (tijdelijk) worden aangepast zodat het beter is ingesteld op die kenmerken.



Tabel 6.2 Schematische weergave van unsupervised adaptation. Van de nieuwe data (audio & tekst) wordt een aangepast akoestisch model gemaakt. Dit model wordt alleen gebruikt om nogmaals dezelfde audio te herkennen.

Keuze: adaptatie

Voor dit afstudeeronderzoek is besloten experimenten uit te voeren met adaptatie. De huidige modellen zijn op vrij grote hoeveelheid data getraind. Hertrainen zou inhouden dat een groot

aantal nieuwe herkenningen toegevoegd moet worden aan de trainingsset, omdat er anders zeer weinig merkbaar is van de veranderingen. Daarnaast is het de vraag of het de prestatie verhoogt, aangezien het huidige systeem is getraind met data waarvan de handmatige transcriptie beschikbaar is, en de nieuwe data minder goed is. De unsupervised adaptatie die gekozen is, is daarentegen op één journaal gericht. Hierdoor is er gelijk resultaat te zien. Het is daarnaast ook waarschijnlijker dat er verbetering in de herkenning optreedt, aangezien de beste data gebruikt wordt.

6.3. *Experimenten*

In dit hoofdstuk zijn de experimenten beschreven die een voorlopig beste adaptatiemethode moeten opleveren. Het doel is grenswaarden te kunnen geven, die de selectiecriteria voor adaptatie zijn. Deze grenswaarden geven aan welke waarde de interne confidence minimaal moet hebben, en wat de maximale WER volgens ondertiteling-REF mag zijn. Om de akoestische modellen beter te maken moet de selectie een bepaalde kwaliteit hebben. Daarnaast moet de selectie ook een bepaalde grootte hebben, anders is er te weinig adaptatiedata. Dan zullen sommige fonemen te weinig voorkomen, waardoor de foneemmodellen slecht geadapteerd worden. Dit kan resulteren in geen verbetering, of zelfs verslechtering van de herkenning.

6.3.1. Opzet

Er is besloten te experimenteren met unsupervised adoptatie, door dataselecties te maken die de beste data bevatten, volgens de monitoring. De adaptatie is gedaan met vijf categorieën die met de hand zijn gekozen. Daarnaast is er nog één controle-categorie die alle data (hele HYP dus) gebruikt voor adaptatie.

De vijf selectiecategorieën zijn op basis van enkele criteria van de monitoring. Er is een categorie met de beste data volgens interne confidencewaarden. Er zijn twee categorieën met de beste (en bijna beste) data volgens scoring met ondertiteling-REF. De vierde categorie is een combinatie van “goed volgens CM” en “goed volgens ondertiteling”. De laatste is een ander soort categorie. Hierbij worden zinsdelen toegelaten die helemaal in de ondertiteling voorkomen. Zo’n serie van (minstens drie) woorden is waarschijnlijk helemaal goed herkend, en zal een lage WER hebben. Deze categorie wordt “100%-regio’s” genoemd.

Om de categorieën beredeneerd te kiezen zijn er experimenten gedaan met SPSS. Deze moeten garanderen dat de data-selecties de juiste zijn, aan de hand van de handmatige transcriptie. In bijlage A zijn alle scripts beschreven die gebruikt zijn voor de verschillende categorieën.

6.3.2. Categorieën kiezen (tuning)

Het doel is in de vier “normale” categorieën 30% tot 40% van de beste data te selecteren. De categorieën zijn: “CM”, “OT-goed”, “OT-veel”, “CM&OT” en “100%-regios’s”. Er wordt dus iets meer aandacht gegeven aan OT-REF-gebaseerde categorieën. Dit is omdat de correlatiecoëfficiënt van de monitoring iets hoger ligt dan die van de interne confidencewaarden. Deze “tuning”-experimenten zijn gedaan op het journaal van 12 jan ’02. Dit journaal is bij zowel CM als ondertiteling het journaal waarop de experimenten zijn gedaan. De tuning-experimenten zijn voornamelijk met SPSS gedaan. De figuren zijn geëxporteerd uit SPSS-analyse-output.

Betrouwbaarheid tuning

De hoeveelheid journaals waarmee gewerkt wordt is klein. Er is voor gekozen de tuning te doen op slechts één journaal, waarna een testset van 2 journaals de antwoorden op de onderzoeksvragen moet geven. Het is mogelijk dat de tuning slechte grenswaarden oplevert omdat het journaal een gemiddelde (of lokale) WER kan hebben die toevallig erg hoog of laag is.

Daarnaast kunnen de confidence-waarden voor dit journaal toevallig bepaalde waarden hebben die in het algemeen niet zo hoog of laag zijn. Naast het kiezen van de categorieën, kan de adaptatie gevolgen ondervinden van de specifieke eigenschappen van dit journaal. Om deze redenen zijn de dataselecties divers, in de hoop dat er ondanks de net genoemde risico's toch een indicatie is te geven van selectie-strategieën voor unsupervised adaptatie.

Interne confidencewaarden

Er is besloten voor deze categorie de beste 35% herkenningen te nemen. Van de 116 herkenningen in totaal zijn dat er dus 41. De beste 41 worden bepaald door de 41 ID's met de hoogste CM te nemen. De 41ste zin heeft als genormaliseerde confidence "-0.07". Als deze waarde als confidence-grenswaarde gekozen wordt, is de categorie "-0.07 of hoger" ontstaan. Daarin bevinden zich 35% van de herkenningen, volgens het idee de beste 35%.

Een kleine analyse op die herkenning: In tabel 6.3 is een SPSS-output te zien waarin de data is gesplitst op de waarde -0.07. Van de beste 35% is gekeken hóe goed deze selectie is.

Report

hrReal

Conf (Banded)	Mean	N	Std. Deviation
<= -.07	40.68%	75	30.840%
-.06+	28.05%	41	27.292%
Total	36.22%	116	30.131%

Tabel 6.3 Een SPSS-report. De hrReal is de WER van de hyp met als referentie de transcriptie, oftewel de "echte WER". De meetmethode is de originele sclite-methode.

Zoals te zien is, is de gemiddelde WER ongeveer 36% voor alle gevallen. Het blijkt dat de categorie "boven -0.07" gemiddeld 8% lagere WER heeft. De selectie heeft dus qua hoeveelheid 35% van de HYP, en een gemiddelde WER van 28%.

OT-veel en OT-goed

Zoals de namen al doen vermoeden zijn deze categorieën op respectievelijk hoeveelheid en kwaliteit gericht. Qua hoeveelheid is de bedoeling dat de categorie "OT-veel" dubbel zoveel zinnen heeft als de categorie "OT-goed". Het idee is dat de volgende categorie de helft van het aantal zinnen heeft, maar betere kwaliteit. Voor de hoeveelheid is de 40% beste data gekozen. Dit houdt in 46 zinnen. De categorie "OT-goed" heeft dus 20% oftewel 23 zinnen.

Het selecteren van de beste data gebeurt door voor de beste ondertiteling-methode de 46 of 23 zinnen met de hoogste WER_OT te nemen. In de categorie "OT-veel" houdt dat in dat de grenswaarde op 34% OT_WER staat. De **hoeveelheid** 40% data hangt dus samen met maximaal 34% **WER_OT**. Verder wordt nogmaals opgemerkt dat de WER_OT, waarop de grenswaarden worden ingesteld, niet hetzelfde is als de "echte" WER. De WER_OT is meestal hoger dan de "echte" WER, welke onderzocht wordt, en in tabel 6.4 en tabel 6.5 in de kolom "Mean" bepaald wordt.

Report

hrReal

FreqT_n2s (Banded)	Mean	N	Std. Deviation
<= 34%	16.96%	46	14.700%
35%+	48.87%	70	31.023%
Total	36.22%	116	30.131%

Tabel 6.4 gemiddelden-analyse van OT-veel(beste 40%). De gemiddelde WER is 36%, de categorie maximaal 34% WER_OT heeft gemiddelde WER van 17%, en de categorie "boven 34%" heeft gemiddelde WER van 49%.

Report

hrReal

FreqT_n2s (Banded)	Mean	N	Std. Deviation
<= 21%	10.00%	24	10.958%
22%+	43.05%	92	29.806%
Total	36.22%	116	30.131%

Tabel 6.5 gemiddelden-analyse van OT-goed (beste 20%). De gemiddelde WER is 36%, in de categorie maximaal 21% (met 24 zinnen) is de WER 10%, boven 21% heeft een WER van 43%

In tabel 6.4 en tabel 6.5 is de gemiddelde WER-verdeling binnen de verschillende categorieën te zien. De categorie OT-goed heeft iets meer dan 23 zinnen, omdat zin 23 en zin 24 dezelfde WER hebben. Het is goed te zien dat het verschil in kwaliteit en kwantiteit bij deze categorieën aanwezig is. OT-veel heeft 46 zinnen, met een gemiddelde WER van ongeveer 17%. Categorie OT-goed heeft 24 zinnen, met een gemiddelde WER van 10%.

CM&OT

Deze categorie is een vrij strenge categorie. De grenswaarden zijn dezelfde als bij CM en OT-veel. De koppel-categorie OT-veel is gekozen omdat de overlap met OT-goed heel klein is. In tabel 6.6 is te zien dat deze categorie ook op kwaliteit gefocust is.

Report

hrReal

Conf (Banded)	FreqT_n2s (Banded)	Mean	N	Std. Deviation
<= -.07	<= 34%	17.88%	26	14.641%
	35%+	52.78%	49	30.394%
	Total	40.68%	75	30.840%
-.06+	<= 34%	15.75%	20	15.068%
	35%+	39.76%	21	31.284%
	Total	28.05%	41	27.292%

Tabel 6.6 gemiddelden-analyse van CM&OT (beste 35% CM en beste 40% OT-veel). De data is gesplitst op CM-theshold. Binnen de CM-delen is onderverdeeld op OT-veel-grenswaarde. De categorie die beide grenswaarden als eis heeft, is zwart omrand.

De selectie heeft 20 zinnen met een gemiddelde WER van 15.75%. Dit is op zich laag, maar in dezefde figuur is te zien dat de bovenste categorie 17.88% WER heeft, wat nauwelijks meer is. Dit doet vermoeden dat de onderverdeling volgens interne confidence minder effect heeft dan volgens ondertiteling-REF.

100%-regio's

Om 100% goed-regio's te bepalen is de scoring-methode "frequentietellen" gebruikt. De versimpelde weergave van het algoritme is als volgt. Per zin wordt per woord gekeken of het woord in de ondertiteling-REF voorkomt. Is dit zo, en is dit voor de minimaal 2 voorgaande woorden ook zo, dan wordt de hele groep woorden als output gegeven. Tevens worden de tijdmarkeringen van de woorden meegegeven zodat de herkenner precies audio en tekst kan koppelen (niet per zin zoals de andere categorieën). Overigens hoeft het voorkomen van een woord in de OT_REF niet persé te betekenen dat het op de juiste plaats voorkomt, en dus tot de goede regio behoort, maar meestal is dit wel zo. De naam "100%-goed" is dus niet helemaal overeenkomstig met de werkelijkheid.

Er zijn 22 regio's met gemiddeld 5.2 woorden, totaal 115 woorden. Daarnaast zijn er nog 25 stilte-momenten. Deze zijn geen echte woorden, maar helpen wel bij de adaptatie van de andere woorden. De gemiddelde zin-grootte van bijvoorbeeld categorie 1: CM is 7.8. Dat zijn dus 41 * 7.8 = 320 woorden. Omgerekend naar hele zinnen is dat 115/320 * 41 = 15 zinnen.

Samenvatting categorieën

Al met al zijn de vijf adaptatie-selectie-categorieën van verschillende samenstelling, zelfs gebaseerd op verschillende monitoringsmethoden. In tabel 6.7 is de kwaliteit en kwantiteit van elke selectie te zien.

Categorie	Hoeveelheid(vd 116)	WER
1. CM > -0.07	41	28%
2. WER_OT < 34%	46	17%
3. WER_OT < 21%	24	10%
4. CM&OT(1&3)	20	15.8%
5. 100%-regio's	15	0%*

Tabel 6.7 Overzicht adaptatie-selectie-categorieën, met daarin de hoeveelheid en de kwaliteit van elke regio. De cel rechtsonder is voorzien van een sterretje omdat de 0% WER niet helemaal gegarandeerd kan worden. De hoeveelheid van deze categorie is omgerekend omdat er geen hele zinnen in categorie 5 zijn.

Er zijn zinnen in bij elke selectie-categorie voorkomen. Van de 116 zinnen zijn er 49 zinnen die buiten alle categorieën vallen. Tussen categorie 1 en 2 (de grootste categorieën) is er een overlap van 20 (dit is categorie 4). Categorie 3 valt natuurlijk geheel binnen categorie 2. Er zijn 11 zinnen die in alle vier selecties voorkomen.

Bij het ontwerpen van deze tuning-experimenten hebben kwaliteit en kwantiteit centraal gestaan. Hoe meer van beiden, hoe beter de te verwachten verbetering in herkenning, na adaptatie. Het is daarom opmerkelijk dat categorie 2 in beide aspecten beter is dan categorie 1. Dit zou inhouden dat categorie 1 geen grote verbetering teweeg zal kunnen brengen, vergeleken met de andere categorieën. Categorie 4 gebruikt ook interne confidencewaarden, en is qua kwaliteit net iets beter dan categorie 2, maar veel minder zinnen.

Er wordt het meeste verwacht van de categorieën die gebaseerd zijn op voorspelde WER aan de hand van de ondertiteling-REF. Het is moeilijk om iets over de laatste categorie te zeggen, omdat het een hele andere opbouw heeft. De 100%-regio's zijn het best vergelijkbaar met de handmatige transcriptie, maar er is heel weinig data van, vandaar dat het de vraag is hoe goed deze zal werken.

6.3.3. Adaptatie-experimenten

Dit zijn de experimenten die kunnen vertellen of, en in hoeverre de unsupervised adaptatie mogelijk is. Ze zijn natuurlijk in grote mate afhankelijk van de in de vorige paragraaf besproken

selectie-categorieën. Deze experimenten zeggen dus veel over de selectie-methoden. De aanpak van de experimenten is qua volgorde gelijk aan het in tabel 6.2 weergegeven schema.

Opzet

De experimenten worden gedaan op de 20.00 uur-journaals van 26 en 27 januari 2002. Dit journaal bestaat uit respectievelijk 255 en 147 zinnen met een gemiddelde WER van 42,1% en 51.8%. Dit zijn hoge WERs.

Deze herkenningen zijn van mindere kwaliteit dan het journaal waar de tuning op is gedaan, die van 12 jan '02.

Naast de vijf selectiecategorieën zoals in 6.3.2, "Categorieën kiezen (tuning)" zijn besproken, wordt adaptatie gedaan op het gehele journaal, dus met de gehele herkenning als transcriptie. Deze categorie heeft 100% data, maar dus hogere WER dan de selectie-categorieën. Van deze categorie werd aanvankelijk vanwege de hoge WER niet veel verwacht, maar het is nodig om de prestaties van de vijf andere categorieën goed te kunnen beoordelen.

Uitvoering experimenten

Na uitvoer van alle monitorings-scripts zijn de zes selectiecategorieën samengesteld. De hoeveelheid zinnen in elk van de categorieën is in tabel 6.8 te zien.

	20.00 journaal 26-jan 2002:		20.00 journaal 27-jan 2002:	
Categorie	Hoeveelheid(vd 255)	perc.	Hoeveelheid(vd 147)	perc.
0. Basis	255	100.00%	147	100.00%
1. CM > -0.07	81	31.76%	39	26.53%
2. WER_OT < 34%	121	47.45%	59	40.14%
3. WER_OT < 21%	73	28.63%	37	25.17%
4. CM&OT(1&3)	46	18.04%	19	12.93%
5. 100%-regio's	55	21.57%	19	12.93%
6. ALL	255	100.00%	147	100.00%

Tabel 6.8 samenstelling selectie bij de adaptatie-experimenten. Per journaal links de categorie, rechts het aantal zinnen. Bij categorie 5 is het aantal zinsdelen omgerekend naar hele zinnen om beter vergelijkbaar te zijn met de andere categorieën.

Procentueel gezien bevat elke selectie gemiddeld ongeveer evenveel data als dezelfde selectie bij de tuning-experimenten, dus op het journaal van 12 jan '02. Dit betekent dat de grenswaarden waarschijnlijk goed ingesteld zijn. Het valt op dat de categorieën 1 en 4 minder data hebben, en de andere juist meer.

Voor elk van deze categorieën wordt vervolgens een nieuw akoestisch model gemaakt (de adaptatie). Daarna wordt de oorspronkelijke data opnieuw herkend. De prestaties van de nieuwe herkenningen worden in tabel 6.9 weergegeven.

	20.00 journaal 26-jan 2002	20.00 journaal 27-jan 2002
Categorie	WER	WER
0. Basis	42.1%	51.8%
1. CM > -0.07	41.8%	50.3%
2. WER_OT < 34%	41.4%	49.8%
3. WER_OT < 21%	41.5%	49.9%
4. CM&OT(1&3)	42.0%	50.7%
5. 100%-regio's	43.9%	55.7%
6. ALL	41.4%	50.7%

Tabel 6.9 Adaptatieresultaten voor beide journaals. Eerste rij de basisscore, daarna de scores na adaptatie.

Er treedt verbetering in, dus de adaptatie lijkt te werken. Het is echter maar maximaal 2% op een originele herkenning van meer dan 50%. Dat is niet zo heel veel. Categorie 5 levert een verslechtering van 5% op.

6.3.4. Evaluatie experimenten

De uitkomst van deze experimenten kan gezien worden als een indicatie van de kwaliteit-kwantiteit-verhouding die gebruikt zou moeten worden bij toepassing van adaptatie. Omdat er maar vijf selectie-categorieën zijn, en adaptatie op twee journaals is getest, kan er nog niet met zekerheid één beste selectiemethode aangewezen worden.

De resultaten van dit experiment geven aan dat er een duidelijk verband lijkt te zijn tussen hoeveelheid data en prestatieverbetering. Er wordt vermoed dat dit komt door de adaptatieparameters van de herkenner. Veel data zorgt voor veel adaptatie. Een beperkte selectie goede herkenningen zal geen grote verbetering teweeg brengen. Daarnaast valt uit de resultaten op te maken dat de “100%-regios”-selectie slecht werkt Deze selectie is de kleinste, en zou volgens de andere resultaten wellicht een wat kleinere verbetering teweeg moeten brengen, maar er is zelfs verslechtering te zien. Hiervoor is geen directe verklaring. Vervolgonderzoek moet aantonen of de selectiemethode slecht is, of dat het niet werkt vanwege fouten in het systeem.

De beste prestatie geeft de selectie van categorie 2, met veel data. Wat verder nog opvalt is dat adaptatie op HYP zonder selectie, ook werkt. Misschien zijn de fout herkende woorden akoestisch niet zo slecht, en heeft de HYP een eenduidige akoestische structuur (zelfde spreker bijvoorbeeld), waardoor in het adaptatieproces de feature-vectors de goede kant op verplaatst worden, en de modellen toch verbeteren.

In tabel 6.9 is te zien dat de **hoeveelheid** data bij de huidige instellingen van de herkenner meer invloed lijkt te hebben op de prestatieverbetering na adapteren dan **hoge kwaliteit** data. Het is zelfs niet zeker dat de selectiemethoden goed werken, aangezien de prestatieverbeteringen zo klein zijn. Om uitspraken te kunnen doen over verschil in kwaliteit bij dataselectie voor adaptatie, is besloten nog een experiment te doen waarbij adaptatie plaatsvindt na aanbieden van selecties van gelijke grootte. Er is alleen verschil in kwaliteit.

Categorie 3, nogmaals

Het experiment is gedaan op het journaal van 27 jan ’06. Categorie 3 is de strengste selectie (beste 20%) op basis van ondertiteling. Aangezien hier het meest van werd verwacht zou dit de selectie zijn met de beste herkenningen. Er zijn nu vier willekeurige selecties gemaakt met dezelfde grootte als categorie 3, oftewel 37 zinnen. Met deze zinnen is hetzelfde adaptatieexperiment gedaan als hierboven beschreven. In tabel 6.10 zijn de resultaten gegeven. Allereerst is de WER_OT gegeven, daarna de echte WER na adaptatie.

	WER_OT	WER na adaptatie
Hele journaal	51.8%	50.7%
Origineel(categorie 3)	15%	49.9%
Random_selectie_1	46%	51.5%
Random_selectie_2	55%	52.8%
Random_selectie_3	53%	52.7%
Random_selectie_4	48%	51.3%

Tabel 6.10 De vier categorieën, met boveninaan het origineel. De middelste kolom geeft de WER_OT, de rechtse kolom de uiteindelijke score.

Hierin is te zien dat de kwaliteit toch verschil maakt. De “categorie 3”, met de beste selectie van data heeft de grootste verbetering tot gevolg na adaptatie. Een adaptatie op de data van de twee slechtere categorieën heeft zelfs verslechtering van de herkenning tot gevolg.

6.4. *Conclusie adaptatie*

Het onderzoek naar training en adaptatie heeft ten eerste een duidelijke voorkeur voor adaptatie gegeven. Training is slechts nodig als de uitspraak structureel verandert en het huidige akoestisch model niet meer blijkt te voldoen. Dit model is van voldoende niveau, dus hoeft niet hertraind te worden.

Dit onderzoek had verder als doel om te onderzoeken in hoeverre unsupervised adaptatie mogelijk is met gebruik van monitoring om de adaptatiedata te selecteren. Het is gebleken dat adaptatie zeker mogelijk is. Adaptatie met een selectie van de herkenningsdata kan een verlaging van de WER opleveren. Toepassing zal in de praktijk direct verbetering tot gevolg hebben. De selectie-strategie is nog niet geheel duidelijk, maar de eerste experimenten hebben aangegeven dat het belangrijk is een groot gedeelte van de herkenning van het journaal te gebruiken, terwijl de kwaliteit niet het belangrijkste lijkt. Na het tweede experiment met selecties van dezelfde grootte is gebleken dat kwaliteit toch wel een groot verschil maakt voor de mate van adaptatie. Het is moeilijk te zeggen wat voor een directe toepassing het belangrijkste is: kwaliteit of kwantiteit, aangezien de huidige instellingen van het adaptatiesysteem ervoor zorgen dat veel data bijna altijd de beste prestaties oplevert.

De selecties op basis van een lage WER_OT blijken goede adaptatieresultaten te geven. De monitoringsmethode met gebruik van teletekstondertiteling lijkt hiermee bewezen te werken.

6.4.1. Aanbevelingen

De resultaten zijn zeker bevredigend, maar adaptatie na selectie volgens het 100%-regios-idee werkt erg slecht. Hier kan op dit moment geen verklaring voor gevonden worden. Aangezien het wél een interessante categorie is, maar er in dit onderzoek geen tijd meer was om het verder te onderzoeken, zal dit in de toekomst onderzocht moeten worden.

De tweede aanbeveling is het ontwikkelen van een systeem van herhaalde adaptatie. Elk journaal moet na adaptatie opnieuw herkend worden, waarna een selectie gemaakt wordt voor een nieuwe adaptatie. Dit herhalen moet doorgaat totdat de WER_OT niet lager wordt dan bij de vorige iteratie. Gezien de enorme rekentijd die adaptatie van het akoestisch model kost, en de tijd gelimiteerd is (elke dag een nieuw journaal dus 24 uur de tijd), is een nieuw adaptatieonderzoek nodig. Hierin zou een strategie ontwikkeld moeten worden om in zo weinig mogelijk iteraties, de beste modellen voor de herkenning te adapteren. Een nieuw kwaliteit/kwantiteit-onderzoek is daarbij nodig. Tevens moet uitgezocht worden of instellingen van het systeem veranderd kunnen worden zodat kleine selecties goede data ook voor een grote verbetering kunnen zorgen.

7. Algemene conclusies

In dit verslag is een onderzoek naar methoden voor montoring van een herkenner voor lopende spraak gepresenteerd. Er zijn twee confidence-methoden besproken, één op basis van interne likelihoods, de ander op basis van teletekst ondertiteling. Om de methoden te testen in een mogelijke toepassing is adaptatieonderzoek en –experimenten gedaan. Hierbij is de data geclassificeerd in categorieën goed en slecht, op basis van de verwachte Word-Error-Rate (WER).

Interne confidence (likelihoods)

Er zijn twee likelihood-methoden onderzocht, met acht normalisatiemethoden. De beste likelihood-methode is de “relatieve confidence” die berekend wordt door de foneem-likelihoods te delen door de likelihood van het best best-scorende foneem op dat moment. Er moet gezegd worden dat er fouten zijn ontdekt in deze methode. In de andere methode de gemiddelde likelihood niet goed bepaald kan worden vanwege de structuur van de herkenner. Hoewel experimenten hebben aangetoond dat de gekozen methode toepasbaar is voor monitoring, staat het vanwege de foutjes, niet onomstotelijk vast dat deze methode ook echt de beste van de twee is.

De beste normalisatiemethode is de combinatie tijd en fonemen. De (relatieve) foneemscores worden gedeeld door het aantal tijdeenheden, daarna wordt per zin de gemiddelde score per foneem bepaald. Deze score is de interne confidence-measure voor de zin. De interne confidence-measure (CM) wordt hierdoor een getal, in de experimenten ongeveer tussen -5 en 5.

De berekende interne confidence-measure is een goede maat voor de kwaliteit van de herkenning. De correlatiecoëfficiënt tussen de confidence-measure en WER is -0.413 wat een duidelijk, negatief lineair verband aangeeft.

Externe confidence (teletekstondertiteling)

De externe confidcence-measure maakt een schatting van de WER door herkenning te vergelijken met ondertiteling. De ondertiteling wordt aan de herkenningen gekoppeld door de tijdmarkeringen te vergelijken, met een marge van 2 seconden aan begin en eind. Hierdoor worden meerdere ondertitelingen samengevoegd en als referentiezin (OT_REF) gebruikt.

Vervolgens wordt het percentage correcte woorden bepaald door HYP en OT_REF te vergelijken. Deze score (de WER_OT of word-error-rate volgens ondertiteling) ligt tussen de 0% en 100% en heeft een duidelijke, positieve lineaire correlatiecoëfficiënt met de echte WER van 0.620. De WER_OT ligt gemiddeld hoger dan de WER. Dit komt doordat de ondertiteling geen perfecte referentie is. De hoge correlatiecoëfficiënt geeft aan dat de WER_OT qua gedrag lijkt op de WER, waardoor de WER_OT een bruikbare monitoringsvariabele is.

Adaptatie

Voor de toepassing van adaptatie zijn van twee journaals selecties gemaakt waarmee nieuwe akoestische modellen getraind zijn. Het ging om zes selecties, die verschillende CM- en WER_OT-categorieën representeerden. De selectie gebaseerd op een WER_OT van maximaal 34% hebben de meeste verbetering in WER opgeleverd. Het is gebleken dat verschil in kwaliteit van adaptatiedata merkbaar verschil oplevert in de herkenning na adaptatie. De instellingen van de herkenner laten het nog niet toe dat kleine hoeveelheden selectiedata grote veranderingen in het akoestisch model veroorzaken, hierdoor is het kwaliteit/kwantiteit-vraagstuk nog niet opgelost. Wel is met dit experiment aangetoond dat toepassing in de praktijk van interne en/of externe confidence-measures werkt.

Centrale conclusie

De monitoring van herkenningen van het broadcast-news lijkt mogelijk, aangezien de interne- en externe confidence lineaire correlatie hebben met de WER van de herkenningen. Gemiddeld gezien is het dus goed mogelijk de WER te voorspellen. Voor individuele gevallen zoals de voorspelde WER van één zin, is het onverstandig deze waarde te gebruiken, aangezien in sommige gevallen het verschil tussen confidence en WER erg groot is (bijvoorbeeld door een slechte ondertiteling, OOV's, etc.). Voor het hele journaal, of per journaal-onderwerp is de monitoring wél betrouwbaar.

Verder moet opgemerkt worden dat de ondertiteling (zeker bij de oude ondertitelingen met samenvattingen) beperkt is. Er zijn herkende zinnen waar geen ondertiteling van is, die kunnen dus ook niet gemonitord worden volgens de methode van externe confidence.

7.1. *Aanbevelingen*

Bij het bepalen van interne confidence-measures is geconcludeerd dat de beide methoden behoorlijk werken. De methode met de hoogste correlatiecoëfficiënt met de WER is verkozen tot beste, maar volgens de literatuur hebben beide methoden hun eigen gebied waar goed gepresteerd wordt. Het zou onderzocht moeten worden in hoeverre een combinatie van beide methoden nog beter WER kan voorspellen. Daarnaast zou onderzoek gedaan kunnen worden naar andere interne confidence-measures, niet alleen gebaseerd op de akoestische likelihoods. Verder moeten de foutjes in de herkenner opgelost worden. De vervolgexperimenten moeten uitsluitsel geven over de beste methode. Voor interne confidence wordt nu met een voorlopig beste interne confidence gewerkt.

De methode voor externe confidence werkt het beste, toch moet dit beter kunnen. De ondertiteling na 2003 is al beter dan de data waarmee voor dit onderzoek gewerkt is. Dat heeft waarschijnlijk direct een lagere gemiddelde WER_OT tot gevolg, en ook een betere dekking van de audio, waardoor de correlatiecoëfficiënt nog beter zal worden. De alignment-methode "oplijning", waar teletekst aan herkenning wordt gekoppeld volgens de tekst in plaats van de tijdmarkeringen, zal beter mogelijk zijn. Deze methode is voor de huidige experimenten niet uitgewerkt, maar dit moet met nieuwe teletekstondertiteling zeker gedaan worden.

De adaptatie-experimenten hebben aangegeven dat de monitoringsmethoden op globaal niveau goed werken. De adaptatie-experimenten geven aan dat na adaptatie nog verbetering mogelijk is, zodat een volgend onderzoek zou zijn naar de strategie van selecties die optimale adaptatie mogelijk maken. Deze selectie gebeurt natuurlijk nog op basis van monitoring.

De oorspronkelijke opdracht is gericht op de SDR-applicatie van de web-portal van het WFH-instituut. Gezien de positieve bevindingen van monitoring in het onderzoek binnen het *broadcast-news*-systeem, is de aanbeveling om de experimenten ook te doen op de data van het WFH-systeem. Hier is geen teletekstondertiteling, maar wellicht dat de interne confidence-methode goed werkt.

De laatste aanbeveling is om de in de inleiding beschreven mogelijke toepassingen uit te werken en implementeren. Het gaat om de **monitoring-tool** en de **rangschikking**. Eerst moet getuned worden wat het (waarschijnlijk lineaire) verband is tussen (interne of externe) confidence en WER. Daarna kan de monitoringtool direct worden geimplementeerd: bereken voor alle segmenten de geschatte WER naar aanleiding van de confidencewaarden. Voor de rangschikking worden meerdere factoren gebruikt die een score opleveren. Er moet uitgezocht worden hoe belangrijk de WER is hierin.

8. Literatuur

Adnergje, B. (2006). "Ondertiteling." from http://nl.wikipedia.org/wiki/Ondertiteling.

Basak, U. B. (2006, 23 jun '06). "Correlatiecoëfficiënt - wikipedia." from http://nl.wikipedia.org/wiki/Correlatiecoëfficiënt.

Burnett, D. C. and M. Fanty (1996). Rapid Unsupervised Adaptation to Children's Speech on a Connected-Digit Task. ICSLP, Philadelphia, PA, USA.

Couvreur, L., J. M. Boite, et al. (2005). "Confidence Measure Normalization for Robust Selection of ASR Agents". International Conference on Speech and Computer(SPECOM-2005), Patras, Greece.

Damper, R. I., M. A. Tranchant, et al. (1996). "Speech versus Keying in Command and Control:Effect of Concurrent Tasking." International Journal of Human-Computer Studies **46**(3): 337-348.

Dolfing, J. and A. Wendemuth (1998). Combination of Confidence Measures in Isolated Word Recognition. 5th Int.Conference on Spoken Language Processsing (ICSLP), Sydney, Australia.

Giuliani, D. and F. Brugnara (2006). Acoustic Model Adaptation with Multiple Supervisions. TC-STAR Wordshop on Speech-to-Speech Translation, Barcelona, Spain.

Gunawardana, A., H. Hon, et al. (1998). Word-Based Acoustic Confidence Measures For Large-Vocabulary Speech Recognition. ICSLP-98, Sydney, Australia.

Hendrickx, R. (2002). "Wat zegt ie?" Retrieved juni '06, from http://www.taaldatabanken.be/taaldatabanken_master/taalbeleid/watzegt.shtml.

Huang, G., W. Hsu, et al. (2003). Automatic Closed Caption Alignment Based on Speech Recognition Transcripts. New York, United States, Colombia University.

Huijbregts, M. A. H. (2005). Shout. Enschede**:** Automatic Speech Recogniser.

Huijbregts, M. A. H., R. J. F. Ordelman, et al. (2005). a Spoken Document Retrieval Application in the Oral History Domain. International conference Speech and Computer. Patras, Greece**:** 699-702.

Jurafsky, D. and J. H. Martin (2000). Speech and Language Processing. Boulder, Colorado, US, Prentice-Hall.

Mengosoglu, E. and C. Ris (2005) Use of acoustic prior information for confidence measure in ASR (automatic speech recognition) applications. Acoustic research letters online **Volume**, 92-98 DOI:

NVVS, N. V. V. S. (2003). ""Wat vindt u van de volledige ondertiteling van het NOS-Journaal van 20.00 uur?" from http://www.oorakel.nl/ondernu.php.

Pellom, B. (2001). SONIC: The University of Colorado Continuous Speech Recognizer. Boulder, Colorado, US, University of Colorado.

Pinto, J. and R. N. V. Sitaram (2005). Confidence Measures in Speech Recognition based on Probability Distribution of Likelihoods. Interspeech-2005, Lisbon, Portugal.

Rose, R. C. and D. B. Paul (1990). "A HIDDEN MARKOV MODEL BASED KEYWORD RECOGNITION SYSTEM". ICASSP 90, Albuquerque, NM, USA.

Shire, M. L. (2001). Relating Frame Accuracy with Word Error in Hybrid ANN-HMM ASR. Eurospeech 2001 - Scandinavia, Aalborg, Denmark.

SPSS (2003). SPSS 12.0.1 for Windows. Chicago, Illinois, United States, SPSS Inc.: Statistical software.

Steehouder, M. F. e. a. (1999). Leren communiceren. Handboek voor mondelinge en schriftelijke communicatie. Groningen, Wolters-Noordhoff.

Thelen, E. (1996). LONG TERM ON-LINE SPEAKER ADAPTATION FOR LARGE VOCABULARY DICTATION. ICSLP, Philadelphia, PA, USA.

Verschuren, P. and H. Doorewaard (2004). Het ontwerpen van een onderzoek. Utrecht, Lemma.

Williams, G. and S. Renals (1997). Confidence Measures for Hybrid HMM/ANN Speech Recognition. Eurospeech '97, Rhodes, Greece.

9. Bijlage A

In dit hoofdstuk is de praktische uitvoer van het afstudeeronderzoek besproken. Hierin wordt ingegaan op de omgeving waarin gewerkt is, de problemen bij het programmeren van de scripts, uitvoer, etc.

Inhoud:

1. interne confidences bepalen
2. ondertiteling
3. adaptatie

9.1. *Inleiding*

Het afstudeerwerk begon met uitgebreide instructies en achtergrondinformatie van de begeleiders, literatuuronderzoek en bestudering van de werkdata. De volgorde zoals hierboven is aangegeven is naast de volgorde waarin de onderzoeken in deze scriptie beschreven zijn, ook de chronologische volgorde van het maken van de scripts geweest.

9.2. *Interne confidences*

Gedurende de afstudeerperiode is het formaat van de *shout*-uitvoer vele malen veranderd. Het bestand “Cmkoppel.pl” moest oorspronkelijk het HYP-bestand met SClite-outputbestanden(PRF) en error-type- en error-region-bestanden koppelen waarna alle fouten per categorie werden uitgespuwd, voorzien van interne confidence.

De laatste versie koppelt geen bestanden meer, maar uit nostalgisch oogpunt is de naam cmKoppel.pl gebleven. Het inputbestand is de originele herkenning inclusief foneemscores etc. Het uitvoerbestand geeft per ID(per zin dus) de interne confidence-waardes, volgens de max-manier en de avg-manier. Zie voor de beide methoden hoofdstuk 4.

```
\>perl –w cmKoppel.pl herkenning.shout > confidences.ljb
```

Het werkt als volgt: bovenin “herkenning.shout” staan de gemiddelde foneemscores. Dit zijn de scores van elk afzonderlijk foneem uit de trainingsset, waarmee dus oorspronkelijk het akoestisch model is getraind. Deze worden opgeslagen. Daarna volgen de zinnen. De zinnen bestaan uit een ID-string, met daarna de fonemen met hun foneem-scores en timeframes.

```
                                Timestamp        TotAM          TotLM          TotComb
#Phone#  SIL m   E              |     0     2 |     3     5 |     6     8 |     -218.75970     -38.91644
#Phone#  m   E   t              |     9     9 |    10    10 |    11    11 |      -22.57382     -13.41707
#Phone#  E   t   SIL            |    12    14 |    15    15 |    16    17 |      -82.22586     -12.68698
                                   17        -323.55939       -60.75762       -384.31702
#Phone#  t   SIL m              |    18    20 |    -1    -1 |    -1    -1 |      -41.49994     -15.76184
                                   20        -365.05933       -60.75762       -425.81696
#Phone#  SIL m   I              |    21    23 |    24    27 |    28    28 |      -69.87055     -11.89417
#Phone#  m   I   n              |    29    29 |    30    31 |    32    33 |      -62.08776     -11.74472
#Phone#  I   n   d              |    34    34 |    35    35 |    36    37 |      -37.96735     -14.78534
#Phone#  n   d   @              |    38    38 |    39    39 |    40    40 |      -49.46277     -13.40265
#Phone#  d   @   r              |    41    41 |    42    50 |    51    59 |     -147.80743     -11.25842
#Phone#  @   r   t              |    60    60 |    61    62 |    63    64 |     -114.78112     -22.68304
                                   64        -847.03632      -149.17963      -996.21594
```

De woorden worden op foneemgrens gescheiden. Per foneem wordt de likelihood in de 1-na laatste kolom opgeslagen. De laatste kolom bevat de max-scores van de foneemloop, oftewel de scores van het beste foneem op dat moment.

Na het inlezen worden de waarden voor de categorieën AVG en MAX uitgerekend (per foneem). Voor AVG is dat “huidige score – avg-score”, voor MAX is dat “huidige score – laatste kolom”. Na elke zin wordt de gekozen normalisatiemethode toegepast. Uiteindelijk wordt dus het bestand **confidences.ljb** opgeleverd met daarin per ID de genormaliseerde AVG- en MAX-scores.

9.3. *Teletekstondertiteling*

In dit hoofdstuk worden de gebruikte scripts voor het bepalen van de juiste teletekstreferenties besproken. Zie hoofdstuk 5 voor het hele onderzoek.

SegOtKoppel.pl

Het eerste script **SegOtKoppel.pl** doet de alignment tussen herkenningen en ondertitelingen. De tijden van de herkenning(of eigenlijk van de ID's) zijn in het dbl- bestand aangegeven.

```
\> perl -w SegOtKoppel.pl journaal.dbl journaal.xml 2000 0 > REF
```

De marge van 2 seconden wordt meegegeven door het 3e argument: 2000(milliseconden). Het laatste is een offset. Als blijkt dat de ondertiteling en dbl structureel een aantal seconden verschillen, doordat bijvoorbeeld de herkenning 4 seconden eerder is begonnen(= dbl-id's tijdstip 0 is tijdstip -4 voor teletekst), dan moet dit argument 4000 zijn.

idMatcher.pl

Heeft als input twee bestanden met ID's, bijvoorbeeld de HYP en REF van een herkenning. Het script moet 2x worden uitgevoerd, met argumenten omgekeerd, om in beide bestanden alleen de ID's te laten staan die ook in de ander voorkomen.

```
\>perl -w idMatcher.pl REF HYP > REF2
\>perl -w idMatcher.pl HYP REF > HYP
```

Dit is nodig om HYP en REF correct te kunnen vergelijken, en omdat SClite eist dat van elke HYP-regel een REF-regel aanwezig is.

stopWoordWisser.pl

Oorspronkelijk geschreven om stopwoorden uit een HYP of REF te verwijderen. De stopwoorden staan in het bestand "stoplist_tno.list". Hierin bevinden zich alle woorden die volgens de samensteller TNO makkelijk worden verwisseld in de allerdaagse spraak. Het is een lijst van 1351 woorden, waarvan een gedeelte hieronder te zien is.

```
1082 vaak
1083 van
1084 vanachter
1085 vanaf
1086 vanavond
1087 vanbinnen
1088 vanboven
1089 vanbuiten
1090 vandaag
1091 vandaan
```

In de huidige experimenten wordt niets gedaan met de stopwoorden. Het idee was deze allemaal te verwijderen zodat de WER van alleen belangrijke woorden zou worden bepaald. Momenteel wordt het gebruikt om silences en leestekens te verwijderen uit HYP of REF.

```
\>perl -w stopWoordWisser.pl HYP 101 > HYP2
```

Het argument 101 staat voor: WEL silences, GEEN stopwoorden, WEL leestekens.

replId.pl

"replace Id". Dit script wordt aangeroepen na het script **basenormpipe.sh.** Dit script is een normalisatiescript, dat bestanden omzet naar een correct formaat met bijvoorbeeld alle getallen uitgeschreven. Voor correcte vergelijking van HYP en REF worden beide bestanden door dit script bewerkt. Na het bewerken zijn de ID's uit de HYP danwel REF. Het script **replId.pl** zet de id's terug.

```
\>perl -w replId.pl oude_HYP nieuw_HYP > nieuw_HYP_ID
```

De "oude_hyp" is een nog niet genormaliseerde versie van de HYP, waarin de ID's nog staan. De "nieuw_hyp" is het bestand na bewerking door basenormpipe.sh.

9.3.1. Scores berekenen

De volgende scriptbestanden worden gebruikt om scorebestanden op te leveren.

keyMapper.pl

Dit script was oorspronkelijk bedoeld om HYP en REF met alleen keywords(dus zonder stopwoorden) te vergelijken. Momenteel wordt daar geen gebruik meer van gemaakt, maar de naamgeving is nog het zelfde. De huidige werking van het script is het scoren van de herkenning volgens de methode "frequentietellen".

```
\>perl -w keyMapper.pl last_hyp last_ref > sc_WER.ljb
```

De argumenten last_hyp en last_ref duiden er op dat alle normalisaties zijn geweest, silences en leestekens zijn verwijderd etcetera. Het uitvoerbestand bevat van elk herkend ID, de Word-Error-Rate(WER).

9.4. *Adaptatie*

De adaptatie maakt gebruik van de gegevens na monitoring. Voor het adaptatie-algoritme worden het originele teletekst-ondertiteling-bestand en de herkenning vereist. De monitoring-scripts worden gebruikt om een confidence-bestand en een ondertiteling-WER-bestand op te leveren. De twee adaptatiescripts zorgen er dan voor dat de gewenste categorieën 1 t/m 5 (zie paragraaf 6.3) bepaald worden.

output_seq.pl

Dit bestand verzorgt de output van de eerste vier categorieën. De zes argumenten bestaan uit bestanden en codes. De bestanden zijn de beide score-bestanden confidences.ljb en sc_WER.ljb, en daarnaast de originele herkenning. De eerste code geeft het type threshold aan. 0 betekent alleen interne confidence(categorie 1), 1 betekent alleen ondertiteling, 2 betekent beide. De tweede code geeft een confidence-threshold aan(hoeft dus bij code 1 niet gebruikt te worden). De derde code geeft de ondertiteling-WER-threshold aan.

```
\> perl -w output_seq.pl confidences.ljb sc_WER.ljb herkenning.shout
0 -0.07 0 > out_1.dbl
```

Alle herkenningen waarvan de waarde in "confidences.ljb" hoger of gelijk zijn aan -0.07, worden geselecteerd.

```
\> perl -w output_seq.pl confidences.ljb sc_WER.ljb herkenning.shout
1 0 34 > out2.dbl
```

Alle herkenningen waarvan in sc_WER.ljb de waarde 34% of lager staat, worden geselecteerd.

```
\> perl -w output_seq.pl confidences.ljb sc_WER.ljb herkenning.shout
1 0 21 > out3. dbl
```

Alle herkenningen waarvan in sc_WER.ljb de waarde 21% of lager staat, worden geselecteerd.

```
   \> perl -w output_seq.pl confidences.ljb sc_WER.ljb herkenning.shout
2 -0.07 34 > out4. dbl
```

Alle herkenningen waarvan in sc_WER.ljb de waarde 34% of lager staat, én in “confidences.ljb” de waarde -0.07 of hoger staat, worden geselecteerd.

De uitvoerbestanden zijn van het type dbl. Deze geven een uitvoer waarin per ID begintijd, eindtijd en herkenning(tekst) zijn aangegeven. Samen met de audio kan de herkenner nu adaptatie doen.

op_100.pl

Dit script is bedoeld voor de 100%-regio’s. Per hyp wordt “frequentietellen” toegepast.

REF: DE gevangenen mogen HUN GELOOF uitoefenen EN WORDEN GELUCHT
HYP: ** gevangenen mogen *** GELOVEN uitoefenen ** ****** *******

HYP	HYP#	REF#
gevangenen	1	1
mogen	1	1
geloven	1	0
uitoefenen	1	1

Score WER = 1 - 3/4 = 1/4

Nogmaals dit figuur ter verduidelijking.

Per HYP-woord wordt gekeken of het in de REF voorkomt. De rijtjes van dit figuur worden daarvoor gebruikt. De naam “100%” doet vermoeden dat het cijfer onder “HYP#” even groot zou moeten zijn als onder “#REF”, maar dit is niet de eis. Dit is omdat er met onbetrouwbare REF, namelijk ondertiteling gewerkt wordt. Als binnen een serie van drie woorden, een van de woorden meerdere keren in de hele HYP voorkomt, maar minder in de REF, is niet te zeggen welk woord uit de HYP de juist herkende was. Om hier een beetje soepel mee om te gaan, is de voorwaarde “50%” geintroduceert.

Als een volgorde van minstens drie woorden allemaal in zo’n frequentie-rijtje links voorkomt, en de waarde in de “#REF” minstens de helft van de waarde onder “#HYP” is, wordt de volgorde geselecteerd voor adaptatie.

Dit houdt niet in dat de methode eigenlijk “50%-regio’s” zou moeten heten want aangezien de woorden voorkomen in HYP en REF, én het kandidaat-woord binnen een volgorde van goede woorden voorkomt, is de kansgroot genoeg om aan te nemen dat het woord inderdaad goed herkend is. De 50% regel is er om ongelukkige HYPs te omzeilen.

```
   \> perl -w op_100.pl last_hyp last_ref herkenning.shout > out5.shout
```

De uitvoer is een bestand van het formaat *shout*, oftewel met foneemscores etcetera. Dit is omdat het om deel-zinnen gaat. Met dit formaat kan de herkenner makkelijker trainen/adapteren bij deel-zinnen.

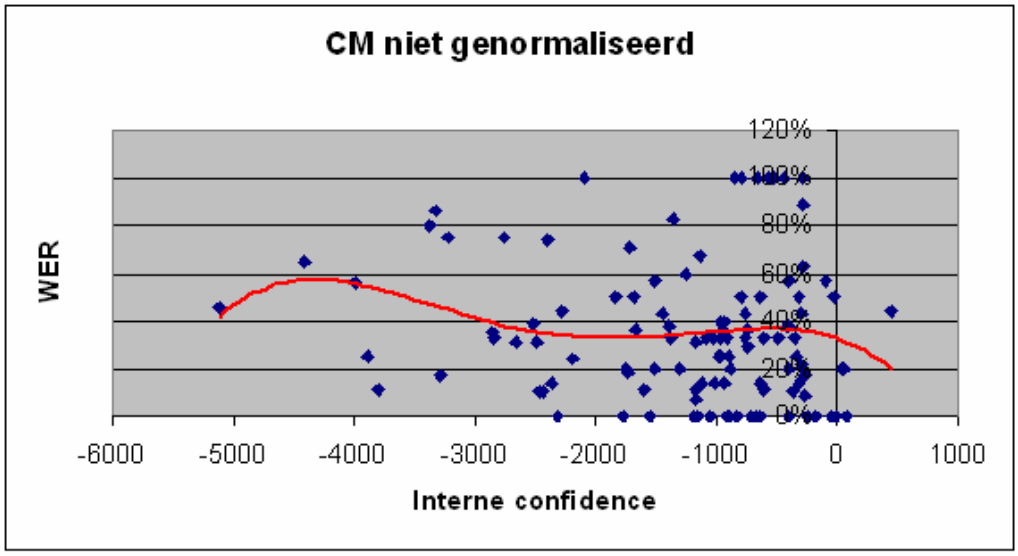
10. Bijlage B

In deze bijlage zijn de resulterende grafieken van alle experimenten in hoofdstuk 4 te zien. Ze zijn in volgorde van normalisatiemethode gedaan, van niet-genormaliseerd tot normalisatie volgens alle drie methoden. Per normalisatie is eerst de correlatiecoëfficiënt en grafiek van de "average"-methode gegeven, daarna die van de "max"-methode.

10.1. *Geen normalisatie*

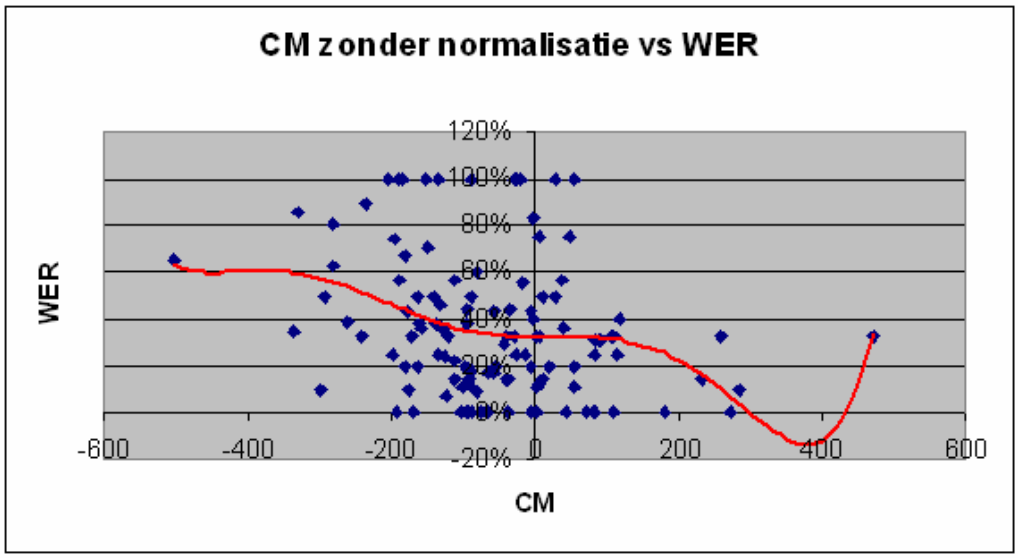
Average

Correlatiecoefficient: -0.094



Max

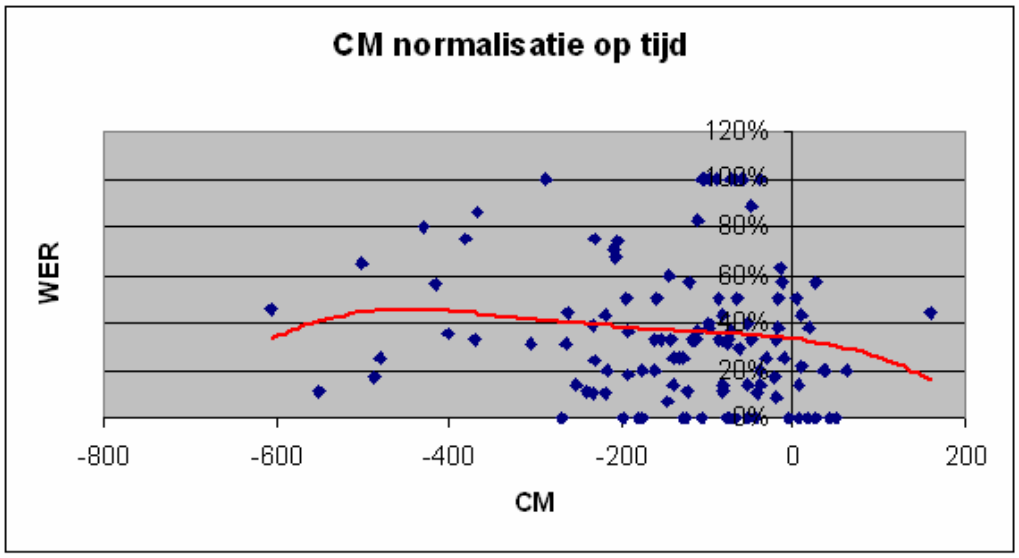
Correlatiecoefficient: -0.264



10.2. *Normalisatie op tijd*

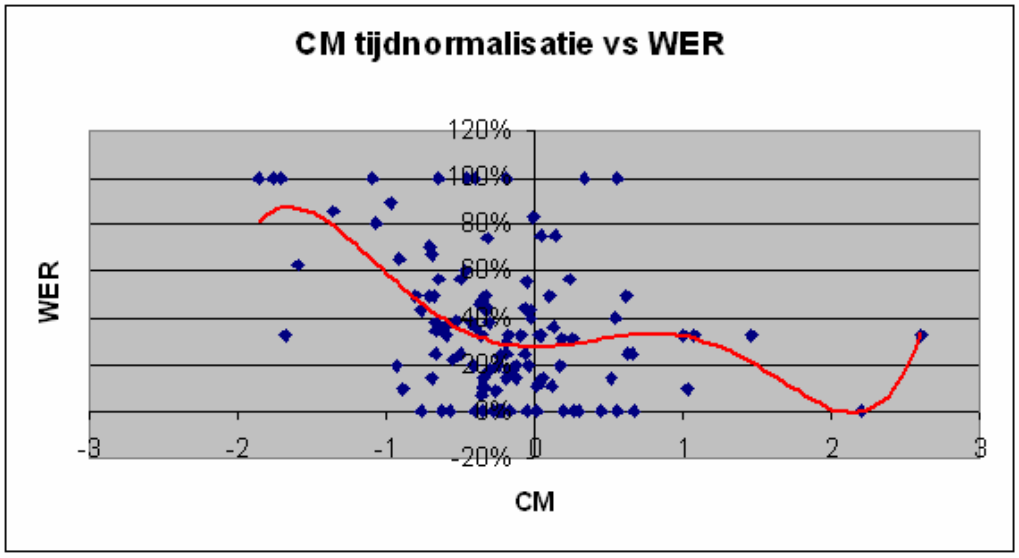
Average

Correlatiecoefficient: -0.332



Max

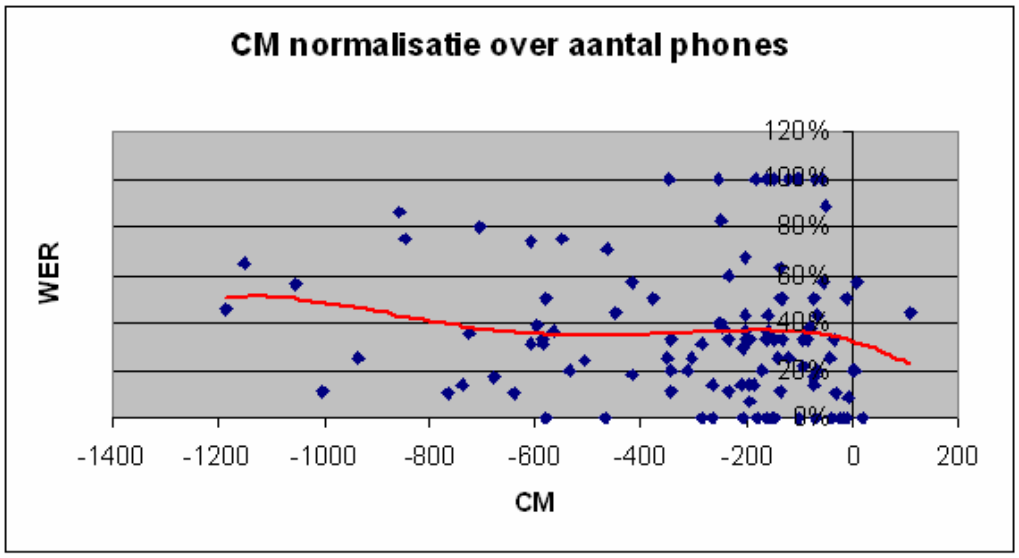
Correlatiecoefficient: -0.355



10.3. *Normalisatie op aantal fonemen*

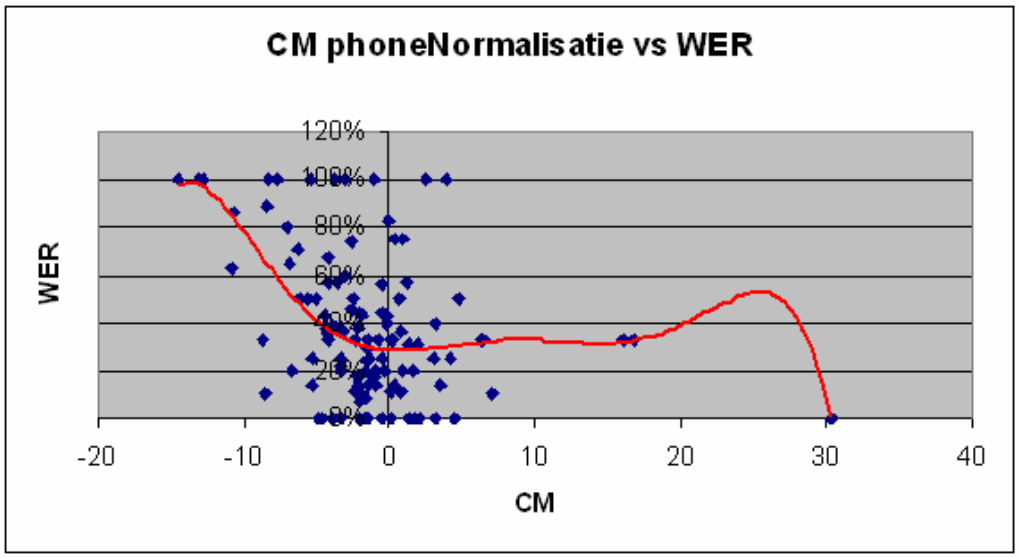
Average

Correlatiecoefficient: -0.335



Max

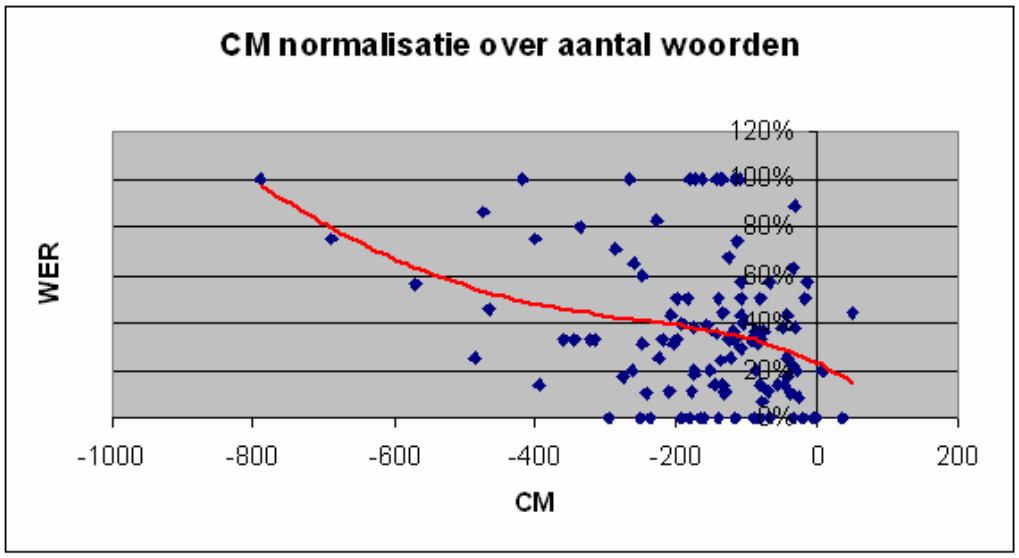
Correlatiecoefficient: -0.354



10.4. *Normalisatie op aantal woorden*

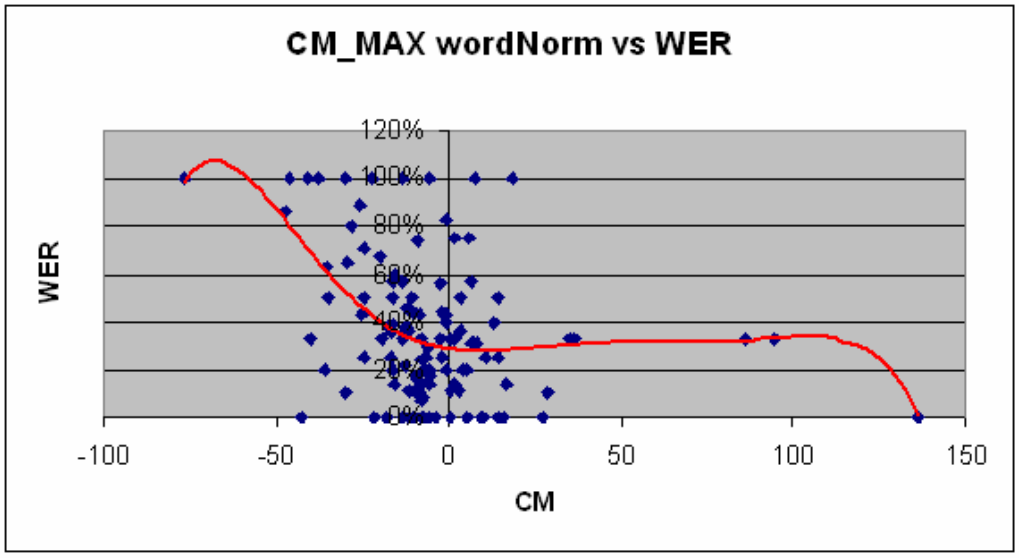
Average

Correlatiecoefficient: -0.323



Max

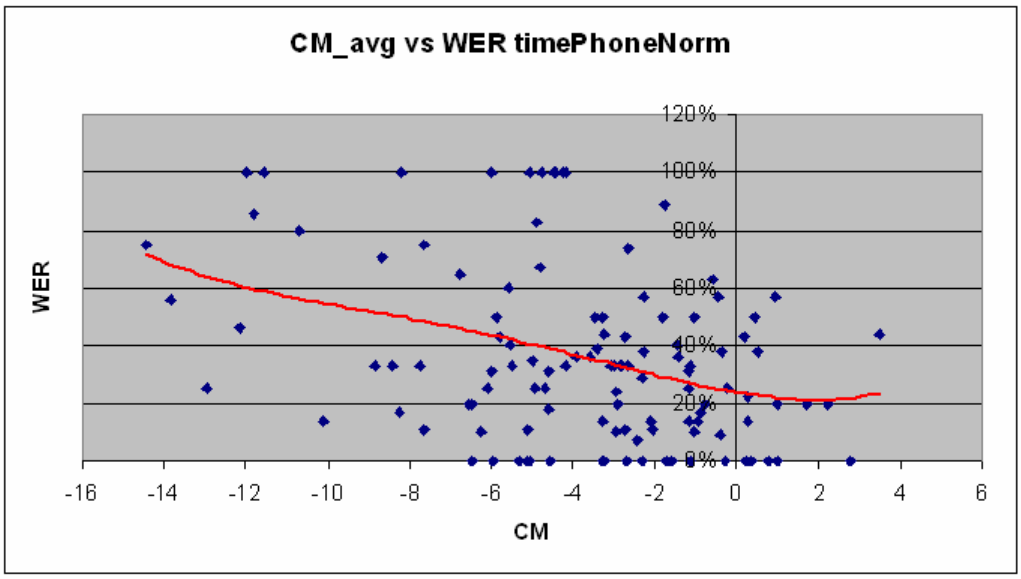
Correlatiecoefficient: -0.319



10.5. *Normalisatie op zowel tijd als aantal fonemen*

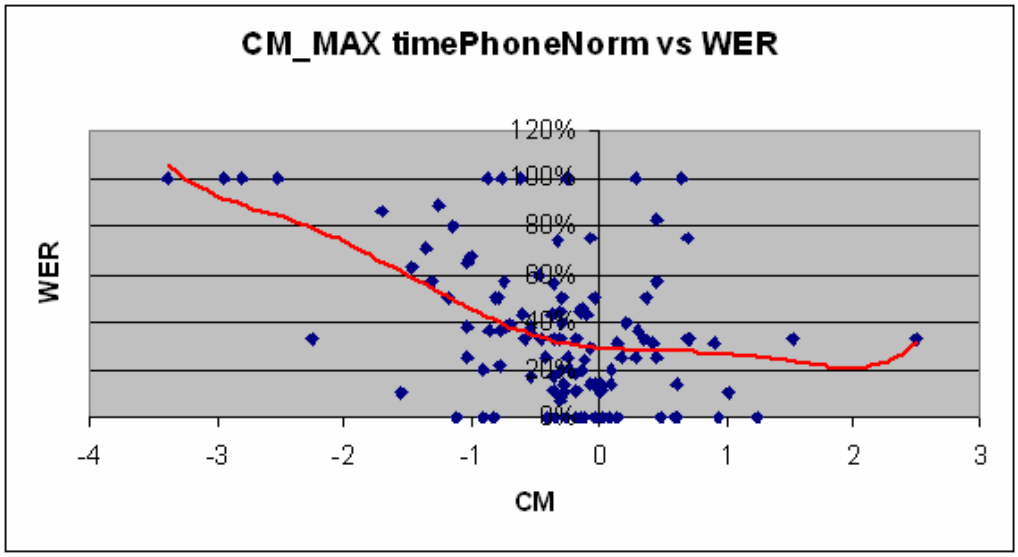
Average

Correlatiecoefficient: -0.362



Max

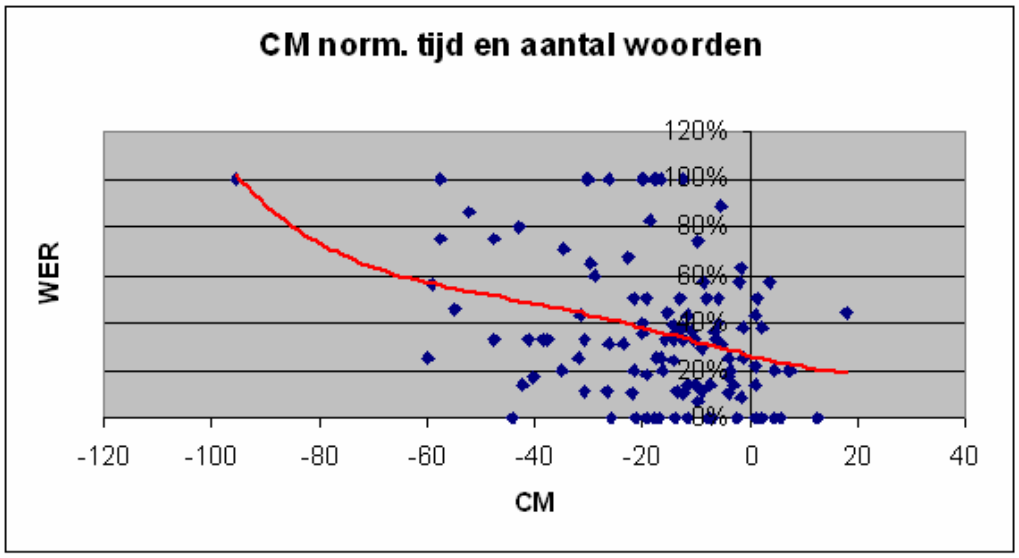
Correlatiecoefficient: -0.413



10.6. *Normalisatie op zowel tijd als aantal woorden*

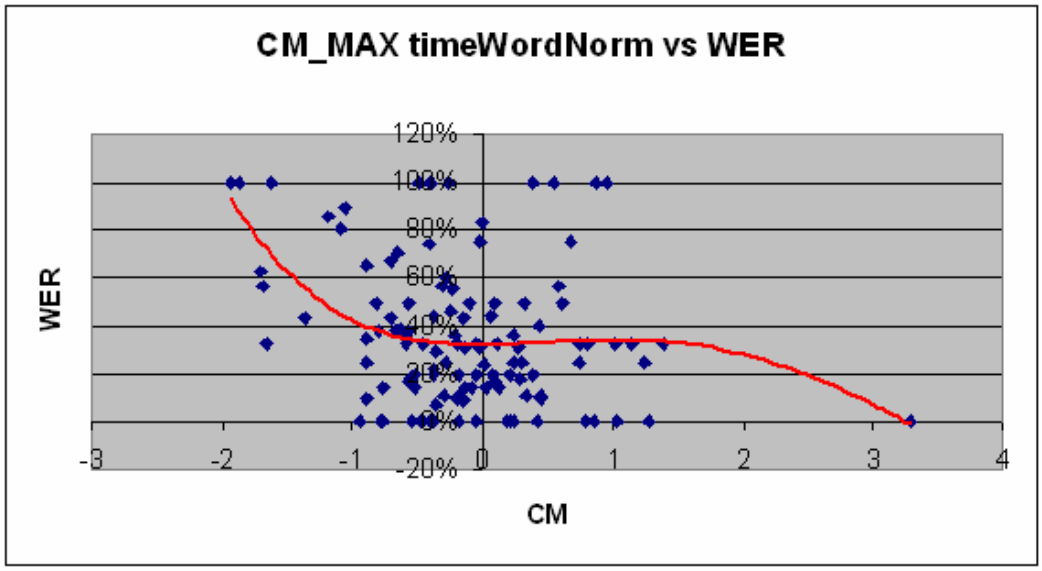
Average

Correlatiecoefficient: -0.341



Max

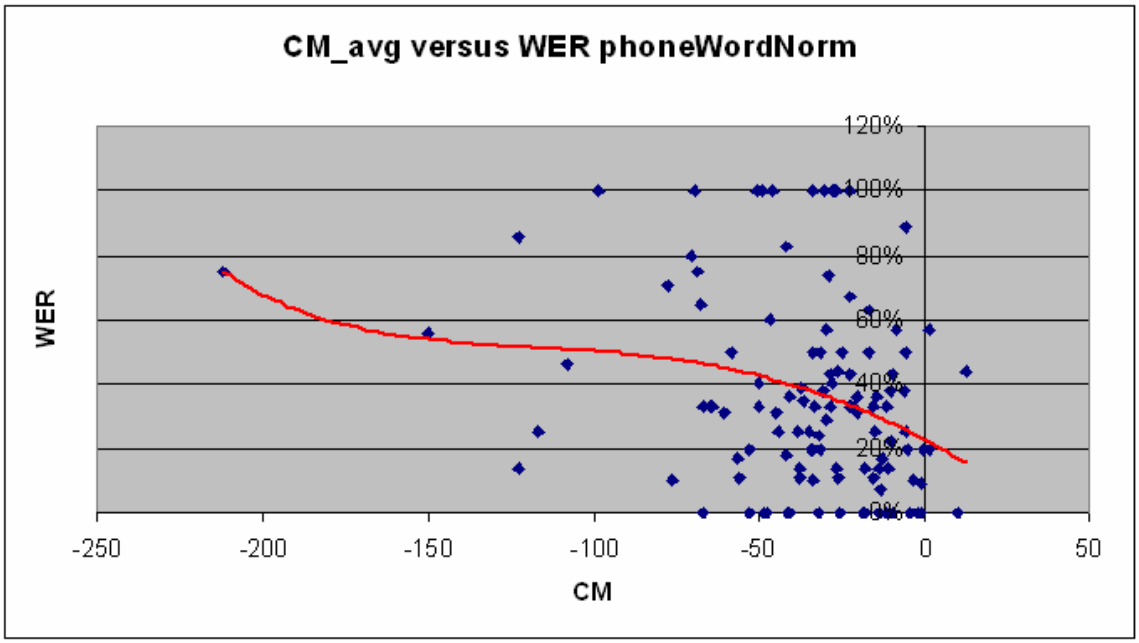
Correlatiecoefficient: -0.260



10.7. *Normalisatie op zowel aantal fonemen als aantal woorden*

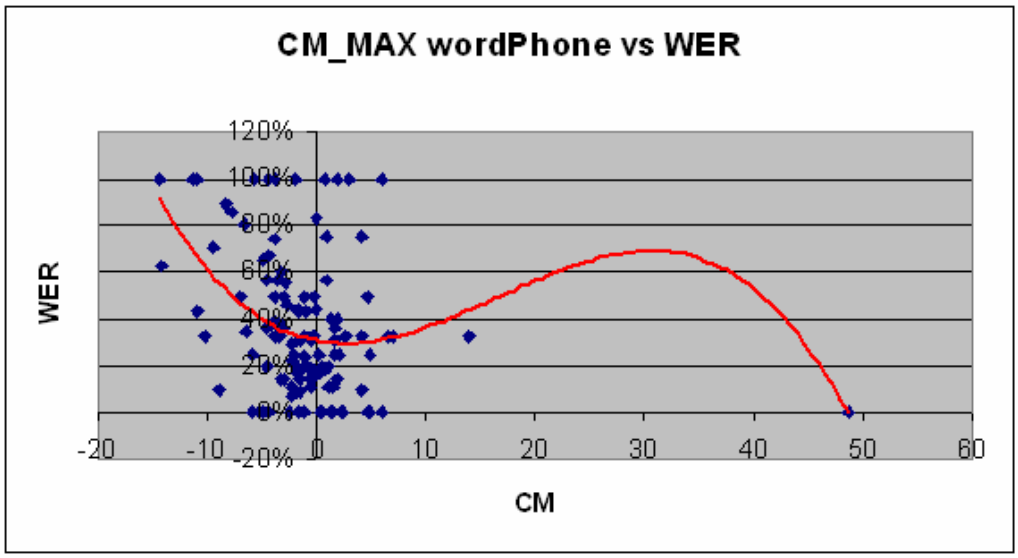
Average

Correlatiecoëfficiënt: -0.272



Max

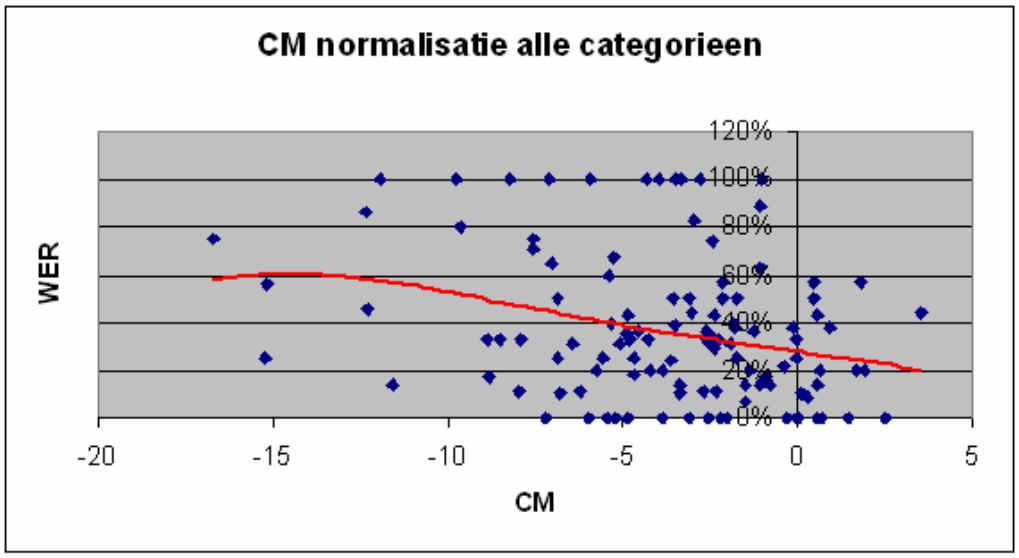
Correlatiecoefficient: -0.272



10.8. *Normalisatie op tijd, aantal fonemen en aantal woorden*

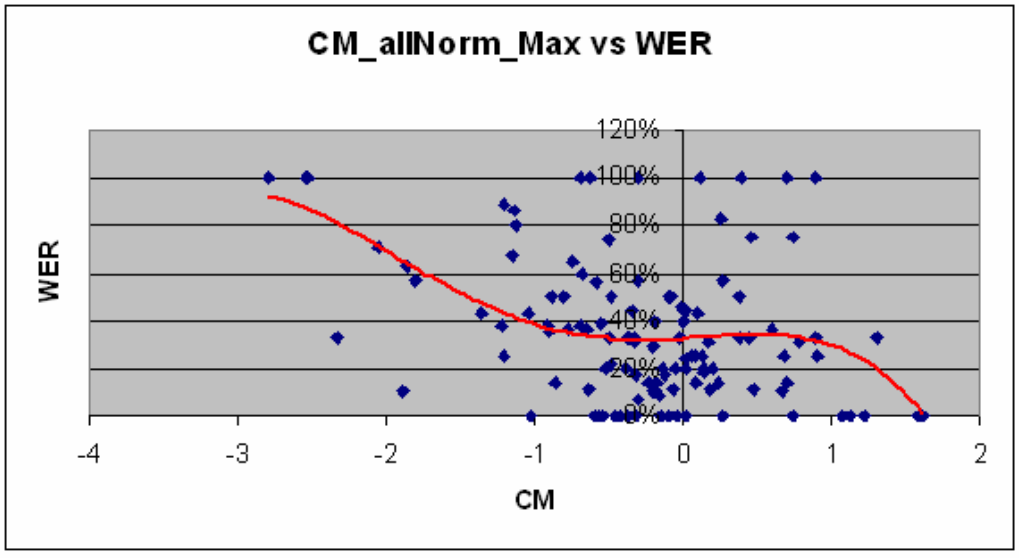
Average

Correlatiecoefficient: -0.288



Max

Correlatiecoefficient: -0.372



11. Bijlage C

Afstudeeropdracht

Titel: On-line monitoring
Student: Laurens Brouwer
Begeleiders: Marijn Huijbregts, Roeland Ordelman, Franciska de Jong
Start datum: 1 september 2005
Eind datum: 1 maart 2006

Omschrijving:

Voor de web-portal van het Willem Frederik Hermans Instituut hebben wij een spraakherkenningssysteem ontwikkeld. Dit systeem herkent de spraak van audio- en videobestanden van interviews met W.F. Hermans en maakt het vervolgens voor de gebruiker mogelijk om in deze bestanden te zoeken. De evaluatie van de performance van dit systeem gebeurt, zoals gebruikelijk bij Automatic Speech Recognition (ASR), door een deel van het automatisch herkende materiaal te vergelijken met een handmatig uitgeschreven herkenning en de zogenaamde Word Error Rate (WER) te berekenen. Deze factor staat voor het percentage woorden dat verkeerd is herkend. Hoe lager dit getal, hoe beter het systeem.

In verloop van tijd zal er steeds meer multimedia data aangeboden worden aan het Willem Frederik Hermans (WFH) systeem. Het zou heel goed kunnen dat deze data steeds minder gaat lijken op de data die initieel gebruikt is om de performance van het systeem te bepalen. Het zou bijvoorbeeld kunnen dat er video van boeksprekingen of praatprogramma's waarin WFH te gast was worden toegevoegd. Deze data zou van een andere kwaliteit kunnen zijn of elementen bevatten waardoor de herkenning slechter wordt (denk bijvoorbeeld aan achtergrond muziek). De WER die eerder gemeten is representeert dan niet meer de daadwerkelijke performance.

Om zeker te weten dat het systeem nog goed draait moet er iets uitgevonden worden om het systeem door de tijd heen, on-line te kunnen blijven volgen zonder dat daar veel handmatig werk bij komt kijken. Met deze opdracht moet hier een eerste aanzet aan gegeven worden. Op dit moment hebben we al een aantal ideeen hoe een monitoringssysteem zou kunnen werken, maar de opdracht zal beginnen met een literatuurstudie waarin bekeken wordt welke technieken al elders zijn onderzocht. Vervolgens worden de mogelijkheden op een rij gezet en moet een monitoringssysteem voor de WFH herkenner geimplementeerd worden. Tenslotte moet een aantal experimenten opgezet en uitgevoerd worden om te meten hoe goed het monitoringssysteem werkt.

Gewenste vaardigheden: C/C++, linux, Perl, of de bereidheid om dit te leren.