



Universiteit Twente
de ondernemende universiteit

Van Mono naar Multi

-

Een automatische VJ

Philippus Baalman

Enschede

Augustus 2007

Voorwoord

Voor u ligt mijn scriptie over mijn afstuderen aan de vakgroep Human Media Interaction van de Universiteit Twente. De opdracht begon met de algemene titel: “Van Mono naar Multi”, vanwege mijn affiniteit met muziek kreeg mijn opdracht daar ook al snel mee te maken en heb ik uiteindelijk een “Automatische VJ” gebouwd.

Allereerst heb ik gekeken of het mogelijk is om een spraakherkenner te gebruiken voor muziek, dit blijkt onder bepaalde voorwaarden goed mogelijk. Daarna heb ik een demonstratie implementatie gebouwd die aan de hand van de songtekst afbeeldingen zoekt en deze op het juiste moment in beeld brengt tijdens het afspelen van een nummer. Vervolgens heb ik deze implementatie voorgelegd aan een groep gebruikers.

Ik wil graag mijn begeleider Roeland Ordeman bedanken die me de nodige zetten heeft gegeven om de scriptie tot een afgerond geheel te maken. Ook mijn andere begeleiders Marijn Huijbregts en Anton Nijholt wil ik graag hartelijk danken voor hun bijdragen aan mijn onderzoek.

Hoewel afstuderen (bij mij in ieder geval) heel vaak kan voelen als een eenzaam proces, is het dat niet geweest. Ik heb tijdens het afstuderen veel steun gekregen van mijn ouders (vanuit Pingjum), mijn zusje Marije (vanuit Berlijn), mijn lieve Carola (vanuit Lima) en mijn appartementgenoot Herman (eerst vanuit Enschede, daarna vanuit Den Haag), waarvoor ik hen uiteraard zeer dankbaar ben.

Summary

This research has the title ‘From Mono to Multi: An automatic VJ’. The goal of this research is to enrich a musical piece (song) with image information. As a guideline for generating that image information the lyrics of a song are suitable. Segmented lyrics on line-level can be used as a basis for search queries in an image database. During the playing of a song those images can be shown.

To be able to show the images at the right moment time information is needed for every line of text. In this research automatic speech recognition is used to generate time labels for lyrics by using forced alignment. An evaluation method has been worked out and applied to a dataset of fifteen Dutch songs.

It can be declared that the forced alignment technique produces good results for use in the automatic VJ program. Nine of the fifteen songs had an alignment of sufficient quality. The six badly performing songs were segmented into separate stanzas and again time labels were generated. This yielded another four alignments of sufficient quality.

An automatic VJ program was made that using the lines of text in the lyrics searches for relevant images in the Google image database. These images are shown to the user at the right moment, according to the time labels.

Finally a short user evaluation was executed from which was shown that people think this combination of images and sound is enjoyable and interesting. Also in general they found that the images fit with the song. Furthermore the user evaluation yielded suggestions for extending the program and further research.

Samenvatting

Dit onderzoek heeft als titel ‘Van Mono naar Multi: Een automatische VJ’. Het doel van dit onderzoek is het verrijken van een muziekstuk (nummer) met beeldinformatie. Als leidraad voor het genereren van die beeldinformatie is de songtekst van een nummer geschikt. Songteksten die gesegmenteerd zijn op tekstregelniveau kunnen gebruikt worden als basis voor zoekopdrachten in een database met afbeeldingen. Tijdens het afspelen van een nummer kunnen die opgezochte afbeeldingen dan getoond worden.

Om de afbeeldingen op het juiste moment te kunnen tonen is er voor elke tekstregel tijdinformatie nodig. In dit onderzoek is automatische spraakherkenning gebruikt om tijdlabels voor songteksten te genereren door middel van forced alignment. Er is een evaluatiemethode uitgewerkt en toegepast op een dataset van vijftien Nederlandstalige nummers.

Er kan gesteld worden dat de forced alignment techniek goede resultaten oplevert voor gebruik in het automatische VJ programma. Negen van de vijftien nummers bleken een alignment te hebben van voldoende kwaliteit. De zes slecht presterende nummers zijn opgedeeld in afzonderlijke coupletten en nogmaals voorzien van tijdlabels. Dit leverde nog eens vier alignments op van voldoende kwaliteit.

Er is een afspeelprogramma gemaakt dat aan de hand van de tekstregels van de songtekst relevante afbeeldingen opzoekt in de Google afbeeldingendatabase. Deze afbeeldingen worden op het juiste moment (aan de hand van de tijdlabels) aan de gebruiker getoond.

Tot slot is er een korte gebruikersevaluatie uitgevoerd waaruit blijkt dat men deze combinatie van afbeeldingen en geluid leuk en interessant vindt. Ook vindt men over het algemeen de afbeeldingen bij het nummer passen. Verder heeft de gebruikersevaluatie bruikbare suggesties opgeleverd voor uitbreiding van het programma en eventueel verder onderzoek.

Inhoudsopgave

1	Introductie	1
1.1	'Van mono naar multi'	1
1.2	Visualisatie van muziek.....	2
1.3	Het onderzoek.....	4
1.4	Outline van het verslag	5
2	Evaluatie Teksten.....	6
2.1	Introductie.....	6
2.2	Automatische spraakherkenning.....	6
2.3	Forced alignment.....	8
2.4	Methode.....	9
2.5	Kwalificatie meetpunten	11
2.6	Beschrijving dataset	12
2.6.1	Categorie I	12
2.6.2	Categorie II	13
2.6.3	Categorie III	13
2.7	Resultaten parametrisering	14
2.8	Resultaten evaluaties.....	16
2.9	Conclusies.....	22
3	Evaluatie beelden.....	23
3.1	Introductie.....	23
3.2	Methode.....	24
3.3	Conclusies.....	27
4	Implementatie	28
4.1	Beschrijving van het afspeelprogramma	28
4.2	Gebruikersevaluatie	29
4.3	Resultaten gebruikersevaluatie.....	30
4.4	Conclusies.....	33
5	Conclusies en aanbevelingen	34
5.1	Conclusies.....	34
5.2	Aanbevelingen	35
6	Referenties.....	37
APPENDIX.....		39
A	Grafieken Categorie I.....	40
B	Grafieken Categorie II.....	42
C	Grafieken Categorie III.....	44
D	Formulier gebruikersevaluatie.....	46

1 Introductie

In dit hoofdstuk wordt de context van de opdracht en de opdracht zelf omschreven. Ook wordt een kort overzicht gegeven van eerdere combinaties van beeld en geluid die er zijn of zijn geweest.

1.1 ‘Van mono naar multi’

Met ‘van mono naar multi’ wordt bedoeld het verrijken van een monomediale stroom met extra informatie. Te denken valt bijvoorbeeld aan de visuele effecten tijdens het afspelen van muziek in populaire muziekafspeelprogramma’s. Een ander voorbeeld is de *Ambient Lighting Technology* van Philips waarbij lichteffecten worden toegevoegd om het televisiescherm heen die corresponderen met wat er op het televisiescherm gebeurt. Dit levert, aldus Philips, een intensere ervaring op tijdens het kijken van televisie [1].

In de toekomst zouden er bijvoorbeeld tijdens een presentatie automatisch relevante beelden achter de spreker kunnen verschijnen. Of een boek zou met text-naar-spraak technologie kunnen veranderen in een voorleesboek.

De focus van dit onderzoek ligt op muziek met gezongen tekst. Instrumentale muziek en muziek waarin alleen geneuried wordt vallen dus buiten dit onderzoek. Als leidraad voor het genereren van beeldinformatie is een songtekst namelijk uitermate geschikt. In songteksten komen personen, objecten, plaatsen, beschrijvingen, ideeën, et cetera voor, die gerelateerd kunnen worden aan een afbeelding. Zo kan de songtekst gebruikt worden als basis voor zoekopdrachten in een database met afbeeldingen. Tijdens het afspelen van een nummer kunnen die opgezochte afbeeldingen dan getoond worden.

Om de relatie tussen getoonde afbeeldingen en de gezongen tekst te kunnen leggen zal de songtekst opgedeeld worden in segmenten. Dit opdelen kan aan de hand van de lay-out van de songtekst gebeuren, bijvoorbeeld op basis van de strofen (coupletten en refreinen), op basis van de tekstregels of per woord. Als bekend is wanneer de segmenten beginnen dan is het mogelijk op het juiste moment (per segment) afbeeldingen te tonen tijdens het afspelen van een muziekstuk. Op dat moment is er van een monomediaal geluidsbestand een multimediaal bestand gemaakt.

Vragen die beantwoord gaan worden zijn:

- Kan een (gesegmenteerde) songtekst automatisch voorzien worden van tijdinformatie?
- Op welke manier kan de koppeling tussen afbeeldingen en de songtekst gemaakt worden?
- Is het mogelijk om passende beeldinformatie toe te voegen aan een muziekstuk?
- Wordt deze toevoeging inderdaad gezien als een verrijking?

1.2 Visualisatie van muziek

Op 1 augustus 1981 begon de televisiezender MTV met 24 uur per dag videoclip uit te zenden, met het toepasselijke getitelde nummer “*Video Killed The Radio Star*” van The Buggles. Sindsdien is een populair popnummer niet meer compleet zonder bijbehorende videoclip. Maar muziek en visualisatie van muziek is al veel vroeger begonnen. Beeldinformatie wordt al eeuwenlang aan geluidsbronnen toegevoegd.

Zo was bijvoorbeeld in 1725 Louis Bertrand Castel bezig met het ontwerpen van zijn *Clavecin pour les yeux* (orgel voor de ogen), waarbij elke noot van het orgel direct gekoppeld was aan de (volgens hem) “juiste” kleur [2].

Bainbridge Bishop publiceerde in 1893 een kleurenschema waarbij noten gekoppeld werden aan de kleuren van de regenboog, hij had toen al drie kleurenorgels geconstrueerd die geluid en corresponderend licht produceren [3], [4].

In de jaren '20 van de vorige eeuw bouwde de Deen Thomas Wilfred een kleurenorgel dat hij de *Clavilux* noemde (zie Afbeelding 1) [5], Hij ontwikkelde wat hij noemde Lumia, de kunst van licht. Een Wilfred Lumia werk is een compositie van licht, kleur en vorm dat tijdens het spelen langzaam verandert [6], [7].



Afbeelding 1: De Clavilux van Thomas Wilfred

Nam June Paik was een Amerikaanse kunstenaar van Zuid-Koreaanse afkomst. Hij wordt gezien als de uitvinder, of in ieder geval de grootvader van de videokunst. Videokunst is genoemd naar de videoband en ontstaan in de jaren '60 van de vorige eeuw. De door Sony in 1965 uitgebrachte Portapak (de eerste draagbare videorecorder) maakte het voor het eerst mogelijk om bewegend beeld op te nemen en direct weer af te spelen. Nam June Paik was in dat jaar de eerste kunstenaar die zo'n Portapak kocht [8], [9]. Eigenlijk was hij daarmee de eerste VJ (Video Jockey of Visuals Jockey).

Ook tijdens concerten worden er regelmatig beelden op de achtergrond (achter de muzikanten) geprojecteerd. Op basis van de muziek worden deze beelden ter plekke samengesteld door een persoon die VJ wordt genoemd. Een VJ manipuleert video op een soortgelijke manier als een DJ (Disc Jockey) zijn of haar platen mixt. Op die manier kunnen bijvoorbeeld (delen uit) nieuwsuitzendingen of films uit hun originele context worden gehaald en hergebruikt worden als beelden in een (multimedia) presentatie of optreden.

Voor het populaire MP3-formaat, waarmee muziek sterk gecomprimeerd kan worden opgeslagen, zijn legio afspeelprogramma's beschikbaar. Bijna al die programma's beschikken over een plug-in die voor visualisatie van het geluid kan zorgen. Geautomatiseerde visualisatie van muziek varieert van eenvoudig tot uitgebreid. Een voorbeeld van een eenvoudige visualisatie is een oscilloscoopdisplay van het geluid. bij een uitgebreide visualisatie worden veranderingen in geluidsterkte, het frequentiespectrum of de baslijn gebruikt om grafische, kaleidoscopische effecten te genereren.

De meeste afspeelprogramma's zoals Winamp, Windows Media Player en Amarok

beschikken ook over de mogelijkheid om de songtekst te tonen tijdens het afspelen van de muziek. Er is echter nog geen afspeelprogramma waarbij de songtekst van muziek wordt verwerkt in de visualisatie van die muziek.

Als er een bron met beeldinformatie beschikbaar is, dan moet het mogelijk zijn om de songtekst en de visualisatie te combineren. Een afspeelprogramma, ofwel een automatische VJ, zou als volgt kunnen werken:

- Door de gebruiker wordt een muziekbestand geselecteerd om af te spelen
- De bijbehorende songtekst wordt door het afspeelprogramma ingeladen
- Aan de hand van de songtekst worden er afbeeldingen uit een database gehaald
- Tijdens het afspelen van het muziekbestand worden op het juiste moment de afbeeldingen getoond, bijvoorbeeld voor elke tekstregel een afbeelding

1.3 Het onderzoek

Om de vragen genoemd in 1.1 te kunnen beantwoorden zal in het onderzoek het volgende moeten gebeuren:

- Er moeten tijdlabls gegeneerd worden voor de songtekst, die aangeven wanneer er wat wordt gezongen om op het juiste moment de afbeeldingen te kunnen tonen
- Er moet een manier gevonden worden om de afbeeldingen aan de songtekst te koppelen
- Er zal een evaluatie moeten plaatsvinden over de afbeeldingen (om te kijken of ze relevant zijn voor de songteksten)
- Er zal een afspeelprogramma geprogrammeerd moeten worden
- Er zal een gebruikersevaluatie plaats moeten vinden om te weten te komen of de toevoeging van de beelden inderdaad gezien wordt als een verrijking

1.4 Outline van het verslag

Hoofdstuk 2 behandelt de evaluatie van de songteksten en gaat dieper in op het gebruik van een spraakherkenner om tijdlabele van een songtekst te genereren. Hoofdstuk 3 gaat vervolgens over de evaluatie van de afbeeldingen, waarna hoofdstuk 4 de implementatie van het afspeelprogramma beschrijft en de gebruikersevaluatie. In hoofdstuk 5 staan de conclusies en suggesties voor eventueel vervolgonderzoek.

2 Evaluatie Teksten

Dit hoofdstuk gaat dieper in op het gebruik van een spraakherkenner om tijdlabels te genereren. Er wordt kort uitgelegd hoe een spraakherkenner werkt. Er wordt een evaluatiemethode beschreven om de kwaliteit van automatisch gegenereerde tijdlabels te beoordelen. Daarna wordt deze evaluatiemethode toegepast op de tijdlabels voor vijftien muzieknummers. Daarna worden de resultaten van die evaluatiemethode besproken.

2.1 Introductie

Omdat voor de gebruiker de relatie tussen beeld en geluid duidelijk moet zijn, zullen de afbeeldingen op het juiste moment op het computerscherm getoond moeten worden. De basis voor de afbeeldingen is de songtekst van een nummer. Een songtekst bestaat alleen uit woorden. Er wordt in de songtekst niet aangegeven wanneer de eerste regel gezongen wordt, evenmin is er iets bekend over de pauzes tussen elke tekstregel. Er moet dus tijdinformatie gegenereerd worden.

Voor dit onderzoek is de tijdinformatie op woordniveau in eerste instantie niet nodig. Tijdinformatie op woordniveau zou wel nodig zijn voor bijvoorbeeld het maken van een automatische karaoke-machine [10].

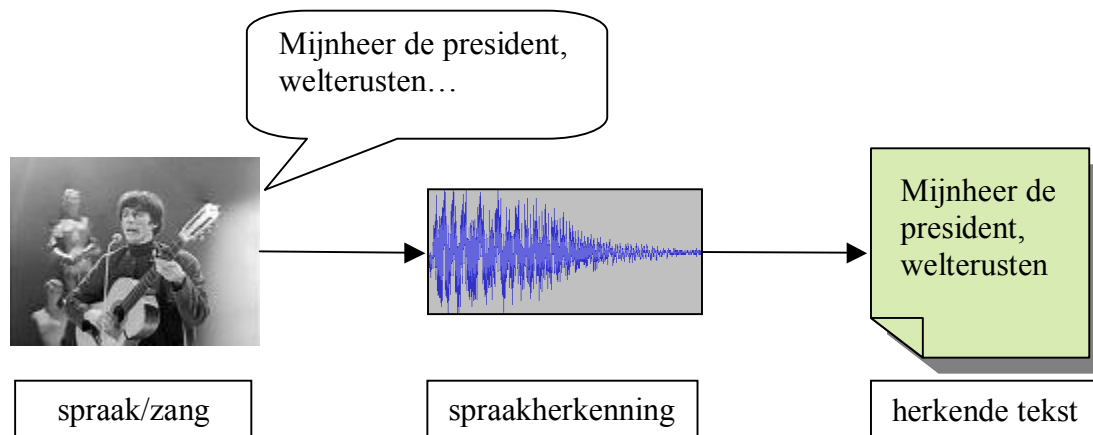
Voor deze implementatie volstaat het om één afbeelding per tekstregel te tonen. Daarom moet er voor elke tekstregel een tijdlabel gegenereerd worden, dat aangeeft wanneer de tekstregel begint. Dankzij die tijdlabels kunnen de afbeeldingen op het juiste moment getoond worden.

Het is theoretisch mogelijk om een spraakherkenner te gebruiken om deze tijdlabels automatisch te genereren. Als door de spraakherkenner voor elk woord in de songtekst een match wordt gevonden in het audiobestand (dit wordt een oplijning of alignment genoemd) dan zijn de tijdlabels bekend.

2.2 Automatische spraakherkenning

Een spraakherkenner zet geluid om naar tekst. De fysieke, meetbare kenmerken van een geluidssignaal zoals frequentie en amplitude (van de geluidsgolf) worden geanalyseerd,

verwerkt en uiteindelijk omgezet naar tekst (zie Figuur 1).



Figuur 1: Schematische weergave van de spraakherkenning. De spraak (in dit onderzoek zang) wordt geanalyseerd door de spraakherkenner en omgezet naar tekst.

Een spraakherkenner maakt gebruik van twee modellen bij de spraakherkenning:

- Het akoestisch model
- Het taalmodel

Het akoestisch model heeft als basis de kleinste onderdelen van de spraak, de fonemen. Woorden, en daarmee de fonemen, kunnen op allerlei verschillende manieren worden uitgesproken. Een mannenstem klinkt bijvoorbeeld anders dan een vrouwenstem. Daarom wordt een akoestisch model getraind met allerlei verschillende uitspraken van de fonemen.

Het andere onderdeel van de spraakherkenner is het taalmodel. Het taalmodel geeft een schatting van de waarschijnlijkheid van een bepaalde woordvolgorde. Een taalmodel moet getraind worden met grote hoeveelheden tekst om die schatting zo precies mogelijk te kunnen geven.

Een spraakherkenner heeft beschikking over een fonetische database die is gevuld met

woorden die herkend kunnen worden en de fonetische spelling van die woorden. Voor woorden die op meerdere manieren uitgesproken kunnen worden staan er ook meerdere transcripties in de fonetische database. De te herkennen woorden worden opgezocht in de fonetische database van de spraakherkenner.

In dit onderzoek wordt een spraakherkenner [11] gebruikt om aan de hand van het muziekbestand en de songtekst een alignment te genereren. Dit kan met een techniek die forced alignment (geforceerde oplijning) wordt genoemd.

2.3 Forced alignment

Forced alignment (geforceerde oplijning) is een techniek die vaak wordt gebruikt om de akoestische modellen van een spraakherkenner te trainen. De spraakherkenner krijgt als invoer een spraaksignaal en een transcriptie. Het systeem kiest dan de beste oplijning tussen het spraaksignaal en de transcriptie. Deze informatie wordt daarna toegevoegd aan het akoestische model.

Deze forced alignment techniek kan in dit onderzoek gebruikt worden om de tijdlabels te genereren. Er wordt dan wel als voorwaarde gesteld dat de songtekst compleet moet zijn en niet meer informatie mag bevatten dan de songtekst. Als er namelijk bijvoorbeeld “refrein” in de songtekst staat zal de spraakherkenner een tijdlabel daarvoor opleveren terwijl het woord “refrein” verder niet in het audiobestand voorkomt.

Het genereren van de tijdlabels zal niet in alle gevallen even succesvol verlopen. Ten eerste zal de aanwezigheid van muziekinstrumenten het proces bemoeilijken en ten tweede verschilt gezongen tekst van gesproken tekst.

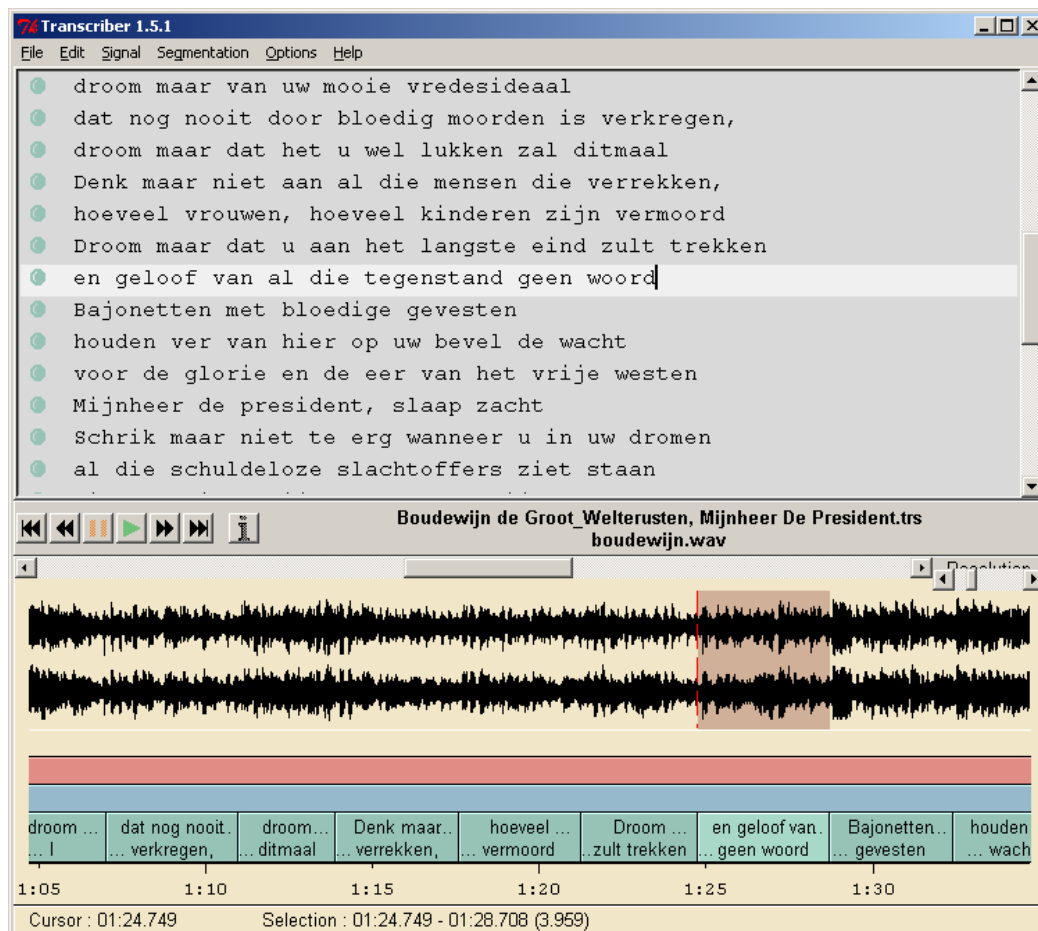
Zo worden klinkers en sommige medeklinkers in zang langer aangehouden dan in gewone spraak. Alle fonemen, vooral klinkers, worden uitgevoerd met meer resonantie. Het volume van zang is hoger, en het toonhoogtebereik is groter dan bij normale spraak. En gezongen fonemen volgen het ritme van de muziek [12].

Om te weten te komen welke omstandigheden er voor zorgen dat een alignment van goede kwaliteit is worden een aantal nummers onder de loep genomen. Deze nummers worden eerst handmatig gesegmenteerd en daarna automatisch opgelijnd met de songtekst. Daarna worden deze oplijningen met elkaar vergeleken.

2.4 Methode

Om te evalueren of een spraakherkenner bruikbaar is voor het genereren van tijdlabels wordt een aantal nummers geselecteerd. Deze nummers fungeren als dataset voor de evaluatie. Van deze nummers is de songtekst overgenomen uit het boekje dat bij de CD zit, opgezocht op internet of met de hand uitgeschreven.

Een annotatieprogramma kan gebruikt worden om handmatig een audiobestand in segmenten in te delen, uit te schrijven en te labelen. Voor elk van de nummers wordt er met behulp van zo'n annotatieprogramma [13] het beginpunt van elke tekstregel van een nummer bepaald en zo handmatig het audiobestand ingedeeld in segmenten (de tekstregels). Omdat het doel is om voor elke tekstregel één afbeelding te tonen is een woord-voor-woord transcriptie niet noodzakelijk en is het voldoende om alleen het beginpunt van een tekstregel van een songtekst te weten. De tijden van deze beginpunten kunnen dan vergeleken worden met de starttijden in de automatisch gegenereerde oplijning die de spraakherkenner oplevert.



Figuur 2: Het annotatieprogramma Transcriber

Tijdens dit handmatige proces komen verschillen tussen de officiële en de gezongen tekst naar voren:

- Soms bevat de songtekst niet alle woorden of is de songtekst bijvoorbeeld gekuist
- Songteksten zijn niet altijd volledig uitgeschreven. Er kan bijvoorbeeld “2x” achter een zin staan die twee keer achter elkaar gezongen wordt. Of het refrein wordt niet opnieuw uitgeschreven, maar men volstaat met een “*herhaal refrein*” tussen de coupletten in
- De songtekst kan typefouten bevatten
- Door een verkeerde interpretatie van de songtekst door de luisteraar die de songtekst heeft uitgeschreven kan het zijn dat deze niet juist is
- De songtekst kan annotaties bevatten die bijvoorbeeld aangeven wat voor soort tekstregels er gaan komen (zoals: “*refrein*”, “*intro*”, “*brug*”), aangeven welke zanger gaat zingen (bij een duet), of geluiden beschrijven (“*handgeklap*”, “*drumsolo*”)

Dit zijn nagenoeg dezelfde zaken als die worden ondervonden bij het oplijnen van meerdere songteksten [14], [15]. Al deze verschillen worden gecorrigeerd zodat er uiteindelijk een transcriptie is met alleen de volledig gezongen tekst en niets anders dan de gezongen tekst. Deze volledige songtekst wordt bewaard voor gebruik in de alignment-tool.

Het taalmodel van de spraakherkenner is getraind met Nederlands gesproken tekst in 16KHz mono formaat. De nummers uit de dataset zullen aangepast moeten worden naar dit zelfde formaat. Daarom worden de ruwe audiobestanden die van de CD afgehaald zijn geconverteerd van stereo naar mono geluid en de frequentie van het muziekbestand wordt ‘gedownsamped’. Nu heeft het muziekbestand het juiste formaat voor gebruik in de alignment-tool. Het muziekbestand en de volledige songtekst zijn de input voor de alignment-tool. Het resultaat is een oplijning op woordniveau met voor elk woord in de songtekst een starttijd en eindtijd. Voor dit onderzoek zijn vooral de starttijden van een tekstregel uit de songtekst van belang.

2.5 Kwalificatie meetpunten

De door de alignment-tool opgeleverde starttijden van een tekstregel worden vergeleken met de ground truth (de handmatig gevonden beginpunten). Dit geeft voor elk meetpunt een afwijking. De gemiddelde absolute afwijking geeft een indicatie van hoe accuraat de automatisch verkregen starttijden van de tekstregels zijn. De afbeeldingen die het afspelprogramma moet laten zien zullen namelijk op tijd in beeld moeten komen (op het moment dat de bijbehorende tekstregels gezongen worden). Het is dus van belang dat de afwijking niet te groot is, omdat dan de afbeeldingen te vroeg of te laat op het beeldscherm verschijnen.

Uiteraard kan het voorkomen dat de uitschieters onder de meetpunten de gemiddelde afwijking te veel beïnvloeden. Er moet dus een andere maat gevonden worden om een alignment goed te kunnen evalueren. Daarom krijgt elk meetpunt ook een score die de kwaliteit van de alignment aangeeft:

1. Zeer bruikbaar (score 1) – het begin van de zin zit in de buurt van het gevonden beginpunt in de handmatige annotatie, binnen een marge van 2 seconden
2. Goed genoeg (score 2) – de alignment wijkt tussen de 2 en 5 seconden af van het handmatig gevonden beginpunt
3. Niet goed (score 3) – de alignment zit niet in de buurt van het handmatig gevonden beginpunt, de afwijking is >5 seconden

Meetpunten met score 1 zijn zeer bruikbaar omdat dan de afbeeldingen op het goede moment getoond kunnen worden. Meetpunten met score 2 worden als ‘goed genoeg’ gekwalificeerd, alhoewel dan de afbeeldingen te vroeg of te laat op het scherm zullen worden getoond. Voor de gebruiker zal de relatie tussen de afbeeldingen en de songtekst echter nog duidelijk zijn. Meetpunten met score 3 zijn eigenlijk onbruikbaar.

Bij een ideale alignment heeft 100% van de meetpunten de score 1, bij een slechte alignment zullen veel meetpunten de score 3 hebben. Om te kunnen spreken van een goede alignment zal een structureel percentage van de meetpunten moeten kloppen met de handmatige transcriptie, ofwel ze moeten score 1 of 2 hebben. In dit onderzoek is dit gezamenlijke percentage op 80% gesteld.

2.6 Beschrijving dataset

Aangezien de trainingsdata voor de spraakherkenner Nederlandstalig is, is er voor dit onderzoek gekozen voor vijftien nummers uit de Nederlandstalige popmuziek. De nummers zijn ingedeeld in drie categorieën, van duidelijk uitgesproken teksten zonder of met minimale begeleiding tot nummers met meer instrumenten en veel bijgeluiden.

Het vermoeden is dat de alignment-tool het beste werkt met nummers uit de eerste categorie (en dat dan dus de alignment van betere kwaliteit is). De eerste categorie bevat nummers waarvan de tekst dichter bij de trainingsdata van de spraakherkenner ligt en de verhouding tussen zang en niet-zang (ruis) gunstig is (als er geluid is, dan wordt er gezongen en wordt er bijvoorbeeld geen gitaarsolo weggegeven).

Voor elke categorie zijn vijf nummers geselecteerd. Hieronder worden bij elke categorie de nummers genoemd. Ook wordt het aantal woorden in de songtekst en de lengte van de nummers genoemd, zodat er later gekeken kan worden of het aantal woorden, de lengte van een nummer of de woorddichtheid van invloed zijn op de kwaliteit van een alignment.

2.6.1 Categorie I

In deze categorie bevinden zich duidelijk gearticuleerde liedjes met minimale begeleiding door instrumenten.

Nummer	aantal woorden	lengte (in min.)	woord-dichtheid
1) Klein Orkest – Over De Muur	271	3:32	1,30
2) Herman van Veen – Dikkertje Dap	192	2:19	1,38
3) Drs. P – Knolraap en lof, schorseneren en prei	335	3:32	1,58
4) Acda en de Munnink – Niet of Nooit Geweest	326	4:03	1,34
5) Raggende Manne – Te Vroeg	171	2:12	1,30

Tabel 1: De nummers in de dataset behorende bij Categorie I, het aantal woorden in de songtekst, de lengte van het nummer en de woorddichtheid (in woorden per seconde).

2.6.2 Categorie II

Dit is de categorie met meer instrumentale begeleiding (gitaren, basgitaren, drums en synthesizers kunnen voorkomen), intro's zonder tekst en minder goed gearticuleerde liedjes. De nummers in deze categorie zouden als 'echte' popmuziek bestempeld kunnen worden.

Nummer	aantal woorden	lengte (in min.)	woord-dichtheid
6) Bløf – Aanzoek Zonder Ringen	303	4:43	1,07
7) Boudewijn de Groot – Welterusten, Mijnheer De President	272	2:36	1,74
8) Spinvis – Voor Ik Vergeet	189	4:25	0,71
9) Het Goede Doel – Zwijgen	314	4:01	1,30
10) Doe Maar – Sinds 1 Dag of Twee (32 Jaar)	229	3:20	1,15

Tabel 2: De nummers in de dataset behorende bij Categorie II, het aantal woorden in de songtekst, de lengte van het nummer en de woorddichtheid (in woorden per seconde).

2.6.3 Categorie III

In deze categorie bevinden zich de nummers met veel bijgeluiden, instrumenten en stemvormingen. De nummers bevinden zich in de genres hip-hop, drum 'n' bass, rock en punk.

Nummer	aantal woorden	lengte (in min.)	woord-dichtheid
11) Bong-Ra – Spiegeltje Met Lijntjes	623	4:15	2,44
12) Opgezwolle - Elektrostress	338	2:08	2,64
13) Heideroosjes – Damclub Hooligan	227	3:18	1,15
14) Di-Rect – Iedereen Is Van De Wereld	230	3:45	1,02
15) The Scene – Blauw	157	4:37	0,57

Tabel 3: De nummers in de dataset behorende bij Categorie III, het aantal woorden in de songtekst, de lengte van het nummer en de woorddichtheid (in woorden per seconde).

2.7 Resultaten parametrisering

Voordat de spraakherkenner in gebruik genomen kan worden voor de evaluatie zal het systeem eerst gecalibreerd moeten worden. Aan de hand van experimenten met het nummer ‘Marco Borsato – Ik Leef Niet Meer Voor Jou’ (dat zou vallen in categorie II) worden de optimale parameters voor de spraakherkenner bepaald die daarna gebruikt gaan worden voor de rest van de experimenten.

De parameters worden één voor één aangepast en na elke aanpassing wordt er een alignment gestart en de score van de alignment bepaald zoals beschreven in sectie 2.4. De resultaten van dit experiment staan in Tabel 4.

Na onderzoek blijkt dat voor de beste resultaten de parameter *-sil* uitgeschakeld moet worden (automatische stiltedetectie en -toevoeging). Er zijn wel een aantal meetpunten met score 1 of 2, maar in het algemeen worden er veel stiltes gedetecteerd die er niet zijn en de gemiddelde afwijking is erg hoog.

De parameter *-con* moet uitgeschakeld worden (context-afhankelijkheid), inschakelen van de parameter levert een hogere gemiddelde afwijking en een lager aantal meetpunten met score 1 en 2.

Voor de parameter *-n* (het aantal actieve toestanden tijdens een zoekproces) geldt, hoe hoger het aantal actieve toestanden hoe beter de kwaliteit van de alignment. Een waarde hoger dan 600 verbetert de resultaten niet meer, maar zorgt wel voor een aanzienlijk langer durende alignment (*-n 600* duurt zo’n 8 minuten, *-n 900* een half uur).

Parameters	Gemiddelde Δ (in sec.)	Score 1 (<2 sec.)		Score 2 (2-5 sec.)		Score 3 (>5 sec.)	
-n 60 -con -sil	19,7	8	20%	6	15%	26	65%
-n 60 -con	7,22	9	23%	8	20%	23	57%
-n 150 -con	14,4	2	5%	1	3%	37	92%
-n 300 -con	19,6	4	10%	0	0%	36	90%
-n 600 -con	13,0	15	38%	3	8%	22	54%
-n 60	23,6	0	0%	0	0%	40	100%
-n 150	20,9	0	0%	0	0%	40	100%
-n 300	14,6	8	20%	4	10%	28	70%
-n 450	14,3	10	25%	4	10%	26	65%
-n 600	10,7	18	45%	7	18%	15	37%
-n 900	10,7	18	45%	7	18%	15	37%

Tabel 4: Resultaten parametrisering ‘Marco Borsato – Ik Leef Niet Meer Voor Jou’. De eerste kolom beschrijft de parameters voor de spraakherkenner, de tweede kolom de gemiddelde absolute afwijking van de alignment ten opzichte van de ground truth. De 3e t/m 5e kolom het aantal meetpunten met score 1, 2 en 3. De vetgedrukte regel laat de parameters voor het beste resultaat zien.

Ter verificatie wordt dit experiment herhaald met een nummer dat in categorie I zou vallen, namelijk ‘Acda en De Munnink – Het Regent Zonnestralen’. Uit Tabel 5 blijkt dat voor dit nummer dezelfde parameters de beste resultaten opleveren (alhoewel bij dit nummer de parameter $-n$ instellen op 300 genoeg is). Deze instellingen worden dan ook gehanteerd voor de andere evaluaties.

Parameters	Gemiddelde Δ (in sec.)	Score 1 (<2 sec.)		Score 2 (2-5 sec.)		Score 3 (>5 sec.)	
-n 60 -con -sil	2,47	20	43%	4	9%	22	48%
-n 60 -con	4,58	14	30%	3	7%	29	63%
-n 150 -con	2,27	33	72%	10	22%	3	6%
-n 300 -con	2,71	31	67%	11	24%	4	9%
-n 600 -con	2,71	31	67%	11	24%	4	9%
-n 60	10,8	18	39%	0	0%	28	61%
-n 150	15,8	18	39%	0	0%	28	61%
-n 300	1,91	42	91%	3	7%	1	2%
-n 450	1,91	42	91%	3	7%	1	2%
-n 600	1,91	42	91%	3	7%	1	2%

Tabel 5: Resultaten parametrisering 'Acda en De Munnink – Het Regent Zonnestrallen'. De eerste kolom beschrijft de parameters voor de spraakherkenner, de tweede kolom de gemiddelde absolute afwijking van de alignment ten opzichte van de ground truth. De 3e t/m 5e kolom het aantal meetpunten met score 1, 2 en 3. De vetgedrukte regel laat de parameters voor het beste resultaat zien.

2.8 Resultaten evaluaties

Met behulp van de parameters die gevonden werden bij de parametrisering worden voor alle nummers in de dataset de score van de alignment bepaald. In Tabel 6 staat voor elk nummer de gemiddelde afwijking vermeld en het aantal meetpunten met score 1, 2 en 3. Ook is het percentage ten opzichte van het totale aantal meetpunten vermeld.

Cat.	Nummer	Gemiddelde Δ (in sec.)	Score 1 (<2 sec.)		Score 2 (2-5 sec.)		Score 3 (>5 sec.)	
I	1) Klein Orkest – Over De Muur	1.72	24	77%	4	13%	3	10%
	2) Herman van Veen – Dikkertje Dap	0.953	29	85%	5	15%	0	0%
	3) Drs. P – Knolraap en lof, schorseneren en prei	0.190	55	95%	3	5%	0	0%
	4) Acda en de Munnink – Niet of Nooit Geweest	1,69	35	73%	10	21%	3	6%
	5) Raggende Manne – Te Vroeg	1,65	27	77%	5	14%	3	9%
II	6) Bløf – Aanzoek Zonder Ringen	5,61	37	67%	8	15%	10	18%
	7) Boudewijn de Groot – Welterusten, Mijnheer De	0,704	34	94%	0	0%	2	6%
	8) Spinvis – Voor Ik Vergeet	33,2	6	23%	2	8%	18	69%
	9) Het Goede Doel – Zwijgen	5,87	21	48%	5	11%	18	41%
	10) Doe Maar – Sinds 1 Dag of Twee (32 Jaar)	14,6	10	29%	8	22%	17	49%
III	11) Bong-Ra – Spiegeltje met Lijntjes	1,56	64	90%	2	3%	5	7%
	12) Opgezwolle – Elektrostress	0,773	49	94%	0	0%	3	6%
	13) Heideroosjes – Damclub Hooligan	7,17	24	56%	1	2%	18	42%
	14) Di-Rect – Iedereen Is Van De Wereld	16,9	11	24%	8	17%	27	59%
	15) The Scene – Blauw	56,0	0	0%	3	8%	36	92%

Tabel 6: Resultaten automatische alignment. Voor alle nummers worden de gemiddelde absolute afwijking, het aantal meetpunten met score 1, 2 en 3 en het percentage ten opzichte van het totale aantal meetpunten vermeld. Bij de roodgekleurde nummers is er sprake van een slechte alignment.

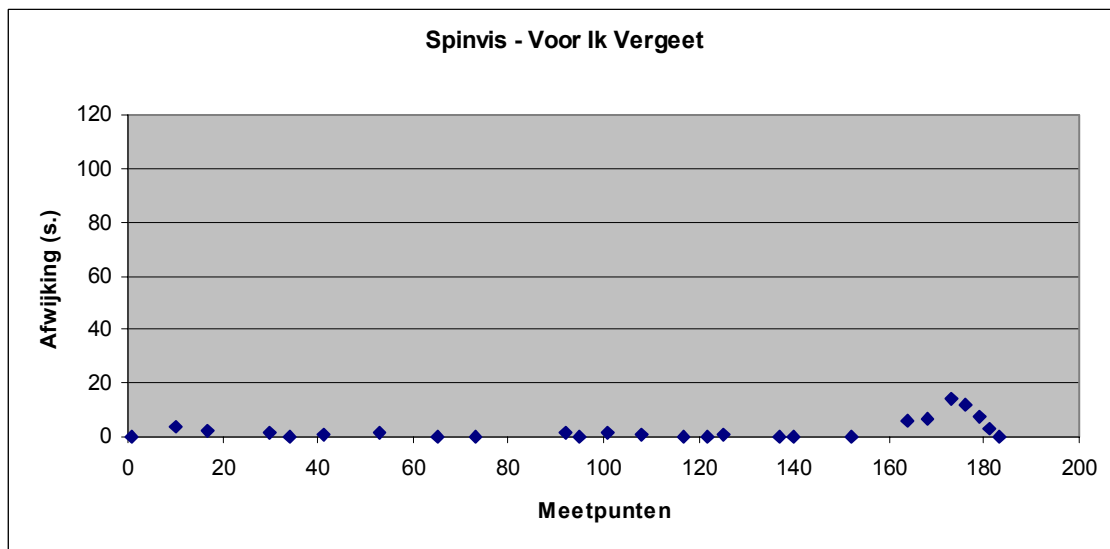
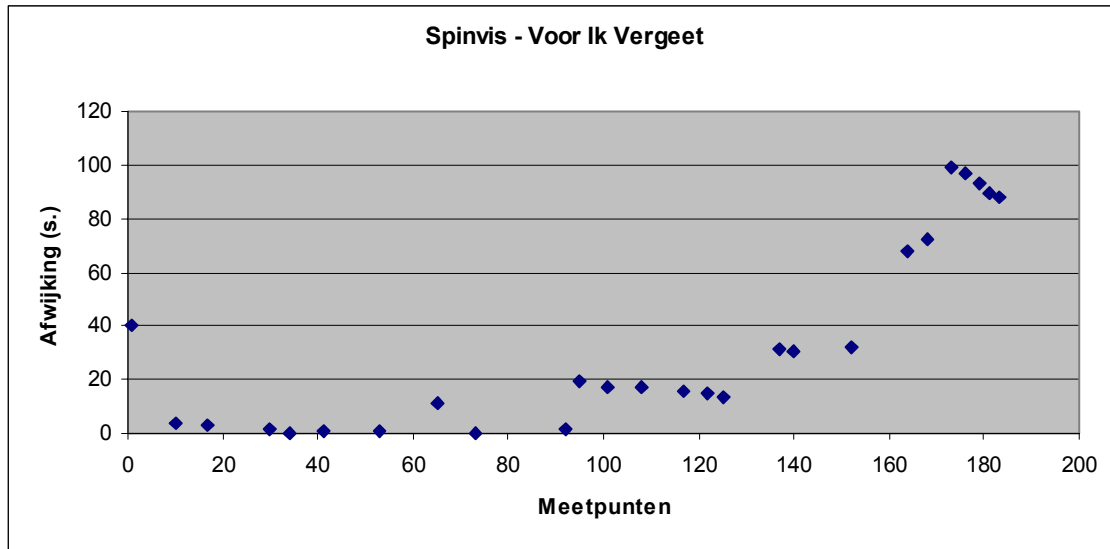
De nummers in categorie I doen het zoals verwacht goed, allemaal leveren ze een goede alignment op. In categorie II leveren twee nummers een goede alignment op en drie nummers niet. Wat vooral opvalt is dat twee van de vijf nummers in categorie III een erg goede alignment opleveren terwijl de verwachting juist was dat de nummers in deze categorie het slecht zouden doen. Deze twee nummers zijn beide hip-hop nummers. In hip-hop is de tekst over het algemeen een zeer belangrijk onderdeel van de muziek. Er komen in een hip-hop nummer dan ook veel woorden voor. Het is de moeite waard om te kijken of de woorddichtheid van een nummer uitmaakt voor de kwaliteit van de alignment.

In Tabel 7 is de lijst met nummers gesorteerd op woorddichtheid weergegeven. Meteen valt op dat de nummers met de slechtste alignment bovenaan staan, en de betere alignments onderaan. Grof gezegd, hoe meer woorden een nummer per minuut bevat hoe beter van kwaliteit de alignment is. Oorzaken voor een lage woorddichtheid zijn bijvoorbeeld een (lang) intro voordat de zanger begint te zingen, instrumentale stukken tussen strofen in of lang aangehouden woorden in de zang.

Nummer	Woord-dichtheid	Kwaliteit alignment
15) The Scene – Blauw	0,567	7,7%
8) Spinvis – Voor Ik Vergeet	0,713	31%
14) Di-Rect – Iedereen Is Van De Wereld	1,02	41%
6) Bløf – Aanzoek Zonder Ringen	1,07	82%
10) Doe Maar – Sinds 1 Dag of Twee (32 Jaar)	1,15	51%
13) Heideroosjes – Damclub Hooligan	1,15	58%
1) Klein Orkest – Over De Muur	1,28	90%
5) Raggende Manne – Te Vroeg	1,30	91%
9) Het Goede Doel – Zwijgen	1,30	59%
4) Acda en de Munnink – Niet of Nooit Geweest	1,34	94%
2) Herman van Veen – Dikkertje Dap	1,38	100%
3) Drs. P – Knolraap en lof, schorseneren en prei	1,58	100%
7) Boudewijn de Groot – Welterusten, Mijnheer De President	1,74	94%
11) Bong-Ra – Spiegeltje met Lijntjes	2,44	93%
12) Opgezwolle – Elektrostress	2,64	94%

Tabel 7: De nummers gesorteerd op woorddichtheid (in woorden per seconde) met de bijbehorende percentages van de meetpunten met score 1 of 2.

Het nummer van Spinvis bevat bijvoorbeeld van de 265 seconden dat het nummer duurt 129 seconden ook daadwerkelijk zang. Er is dus veel ruis wat de oplijning lastig maakt. Een mogelijke oplossing is om het nummer op te delen in coupletten en per couplet op te lijnen (dit wordt vaker gedaan, zie bijvoorbeeld [16]). In Grafiek 1 en 2 is het verschil te zien tussen de oplijning van het complete nummer zonder opdeling en de oplijning na opdeling in coupletten. Het opdelen van het nummer blijkt dus inderdaad de alignment te verbeteren.



Grafiek 1 en 2: Het verschil tussen oplijning van het complete nummer van Spinvis zonder opdeling, daaronder de oplijning van het nummer na opdeling in coupletten.

De grafieken in de appendix doen vermoeden dat meer nummers op deze manier voorzien kunnen worden van een betere oplijning. Als de andere slecht presterende nummers ook worden opgedeeld en opnieuw opgelijnd dan blijken nu alle nummers in categorie II een goede alignment op te leveren en er is één extra nummer in categorie III voorzien van een goede alignment (een overzicht staat hieronder in Tabel 8).

Nummer	voor opdeling	na opdeling
8) Spinvis – Voor Ik Vergeet	31%	81%
9) Het Goede Doel – Zwijgen	57%	93%
10) Doe Maar – Sinds 1 Dag of Twee (32 Jaar)	51%	80%
13) Heideroosjes – Damclub Hooligan	58%	100%
14) Di-Rect – Iedereen Is Van De Wereld	41%	67%
15) The Scene – Blauw	8%	49%

Tabel 8: Het verschil in kwaliteit van de oplijning zonder opdeling en met opdeling in coupletten.

De overige nummers (die al wel voorzien waren van een goede alignment) zouden met deze techniek een nog nauwkeurigere alignment kunnen krijgen. Als de automatische alignment tot op woord-niveau goed werkt, zou het afspeelprogramma uitgebreid kunnen worden tot een karaoke-machine. Om te testen of de automatische alignment op woord-niveau goed is, zouden een aantal nummers handmatig volledig op woordniveau moeten worden gesegmenteerd en gelabeld. Daarna kan geëvalueerd worden of de automatische alignment hiervoor geschikt is.

2.9 Conclusies

Er is in dit hoofdstuk een evaluatiemethode uitgewerkt om de kwaliteit van een alignment te beoordelen. Na parametrisering van de spraakherkenner zijn de alignment-scores voor een vijftiental nummers bepaald.

Voor nummers met een slechte alignment blijkt het te helpen als het nummer wordt opgesplitst in delen en dat daarna die delen apart worden opgelijnd. De woorddichtheid van een nummer geeft een indicatie over hoe succesvol de alignment zal verlopen. De uiteindelijke implementatie zou de gebruiker kunnen waarschuwen als de woorddichtheid laag is.

Er kan gesteld worden dat de spraakherkenner bruikbare resultaten oplevert voor gebruik in het afspeelprogramma. Bij negen van de vijftien onderzochte nummers heeft tenminste 80% van de meetpunten score 1 of 2. Na opdeling in coupletten en opnieuw oplijnen van de slecht presterende nummers zijn dat er nog eens vier. De automatisch gegenereerde tijdlabele zijn dus in veel gevallen geschikt voor gebruik in het afspeelprogramma om op de juiste momenten de afbeeldingen op het scherm te tonen.

3 Evaluatie beelden

Dit hoofdstuk gaat dieper in op de afbeeldingen die nodig zijn voor het afspeelprogramma. Er wordt een aantal mogelijkheden beschreven waarmee afbeeldingen kunnen worden opgezocht. Er wordt een database gekozen als bron voor de afbeeldingen en de zoekopdrachten worden afgesteld zodat er relevante afbeeldingen uitkomen.

3.1 Introductie

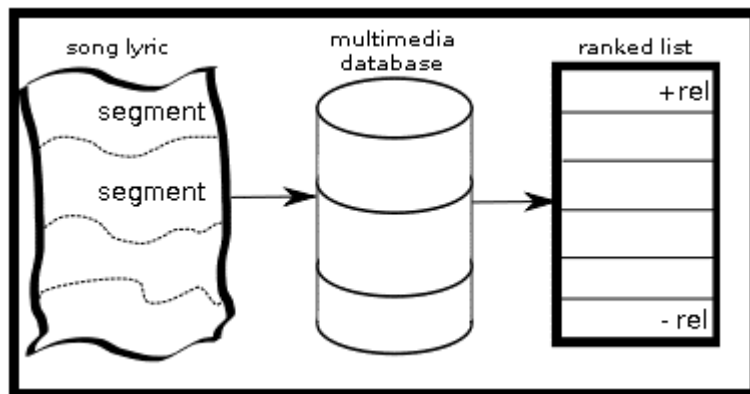
Voor het koppelen van beelden aan de songteksten zijn er verschillende vormen denkbaar. Een afspeelprogramma zou bijvoorbeeld de harde schijf van een gebruiker kunnen gebruiken om zo een collectie van beelden te verzamelen. Dan ziet de gebruiker afbeeldingen uit zijn eigen fotocollectie terug bij het afspelen van de muziek. Nadeel is dat de afbeeldingen dan waarschijnlijk niet goed beschreven zijn. De bestandsnaam van de afbeelding en de naam van de folder zijn dan de enige indicaties over het onderwerp van de afbeelding.

Bij het afspeelprogramma zou een collectie goed beschreven afbeeldingen meegeleverd kunnen worden. Nadeel is dat er altijd songteksten zullen zijn met onderwerpen die niet in de collectie zitten en dan is een online bron voor de afbeeldingen beter geschikt.

Er zijn verschillende online databases gevuld met afbeeldingen zoals *Flickr* [17] en *Zoomr* [18] waar de afbeeldingen aan de hand van tags of labels kunnen worden opgezocht. Veruit de bekendste database met afbeeldingen is *Google Image Search* [19]. In de Google zoekmachine kan aan de hand van het invoeren van een zoekopdracht op afbeeldingen worden gezocht. De sleutelwoorden voor het vinden van afbeeldingen zijn bij de zoekmachine gebaseerd op de bestandsnaam, de hyperlinks die naar het plaatje verwijzen en tekst die in de buurt van het plaatje op de internetpagina wordt genoemd. Google gebruikt algoritmes om duplicaten van afbeeldingen uit de zoekresultaten te filteren en afbeeldingen met de hoogste beeldkwaliteit worden als eerste gepresenteerd. Dit alles maakt *Google Image Search* een uitstekende kandidaat om te gebruiken als bron voor het afspeelprogramma.

3.2 Methode

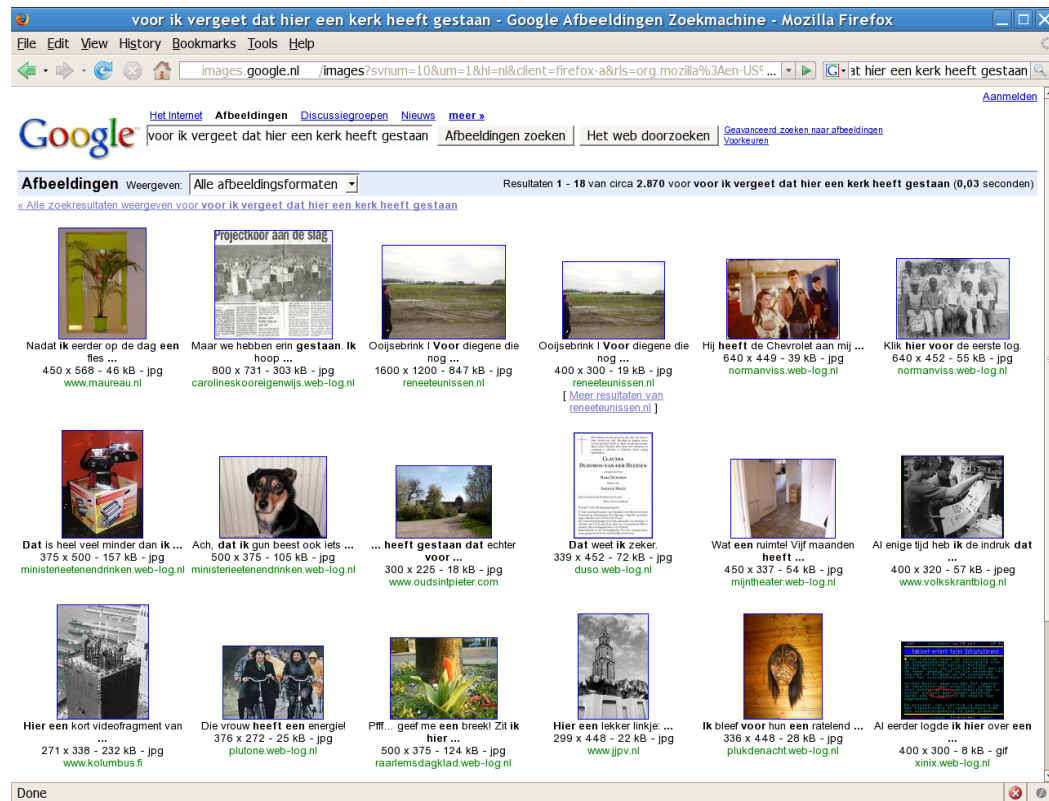
De songtekst is opgedeeld in tekstregels. Elke tekstregel wordt gebruikt als zoekopdracht in Google Image Search. Zo kan een zoekopdracht gezien worden als een zoekopdracht in normale gesproken taal, ofwel een natuurlijke taal query. Het resultaat van de query is een ranked list van van afbeeldingen. In het geval van Google Images zijn de afbeeldingen gesorteerd op relevantie en beeldkwaliteit.



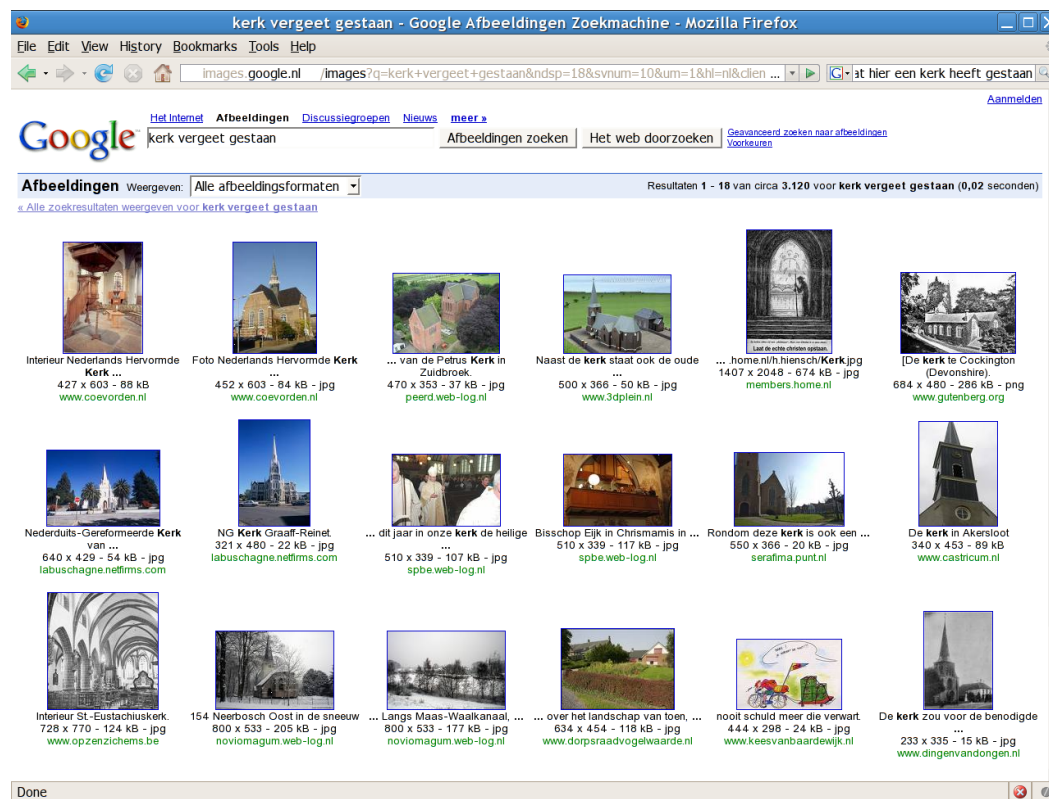
Figuur 3: de gesegmenteerde songtekst is de basis voor een query in een multimedia database. De query levert een ranked list van afbeeldingen op geordend naar relevantie.

Uit een pilot experiment met de tekstregel: “Voor ik vergeet dat hier een kerk heeft gestaan.” blijkt dat de zoekopdrachten weinig relevante plaatjes opleveren (zie Figuur 4). Dit komt doordat de zoekopdracht een aantal in het Nederlands frequent voorkomende woorden bevat zoals: “voor”, “ik”, “dat”, “hier” en “een”. Bij Google Image Search worden deze woorden er niet uitgefilterd. Deze woorden zijn echter weinig relevant voor het vinden van afbeeldingen.

Om het zoekresultaat te verbeteren wordt een zogenaamde stoplijst gebruikt om deze woorden uit de zoekopdracht te filteren zodat deze gereduceerde zoekopdracht overblijft: “kerk vergeet gestaan”. Als deze aangepaste zoekopdracht wordt ingevoerd in *Google Image Search* komen er wel relevante resultaten terug, in dit geval veel afbeeldingen van kerken en vergeet-me-nietjes (zie Figuur 5).



Figuur 4: Voorbeeld van een zoekopdracht in Google Image Search



Figuur 5: De met behulp van de stoplijst gefilterde zoekopdracht

Met hulp van een semi-onafhankelijke vrijwilliger, die met een frisse blik naar de afbeeldingen kon kijken, is er voor één van de nummers (nummer 12 uit de dataset - Opgezwolle – Elektrostress) gekeken of de afbeeldingen enigszins relevant zijn voor de tekstregels. Dit blijkt in de meeste gevallen zo te zijn, er komen echter wel een aantal problemen naar voren:

- Sommige Nederlandse woorden hebben ook hun betekenis in een andere taal. Het gekozen nummer bevat bijvoorbeeld het woord ‘dood’ dat in het Engels straattaal is voor ‘dude’. Om een betere relevantie te krijgen tussen afbeelding en songtekst is er besloten om een extra parameter *site:.nl* toe te voegen aan de zoekopdrachten in Google Image Search. Dit beperkt de zoekopdrachten tot Nederlandse webpagina’s.
- Bij sommige tekstregels kan geen afbeelding gezocht worden omdat alle woorden in de tekstregel in de stoplijst voorkomen. Een mogelijke oplossing is om dan de huidige afbeelding op het scherm te laten staan, een andere oplossing is om een afbeelding te gebruiken uit een eerdere zoekopdracht.
- Een aantal zinnen leveren afbeeldingen op van de band of de platenhoes van een album van die band. Dit kan voorkomen als bijvoorbeeld in een recensie wordt gerefereerd naar de songtekst, of als de woordencombinatie zo uniek is dat ze alleen voorkomt op een webpagina die gerelateerd is aan de band. Dit is op te lossen door woorden die niet mogen voorkomen op de webpagina waar de afbeelding staat uit te sluiten. Dit kan door een min-teken voor een woord te zetten. De parameter *-acda* zou bijvoorbeeld pagina’s uitsluiten waar ‘acda’ in de tekst voorkomt.
- Het kan ook voorkomen dat tekstregels geen afbeeldingen opleveren. Dit is soms op te lossen door met de zoekopdracht niet te zoeken naar **alle** woorden, maar naar **één of meer** woorden. Dit komt neer op het toevoegen van de operator *OR* tussen de zoektermen.

3.3 Conclusies

Er is in dit hoofdstuk een basis gelegd voor het opzoeken van de afbeeldingen aan de hand van een songtekst. Er is een database gekozen als bron voor de afbeeldingen. Bij het testen van de database kwamen een aantal problemen naar voren die door het aanpassen van de queries eenvoudig te verhelpen zijn.

Nu kan er begonnen worden met de implementatie van het afspeelprogramma, deze implementatie wordt beschreven in het volgende hoofdstuk. Bij het afspelen van de nummers zal de muziek weer terugkomen in het onderzoek. Het kan dus zijn dat afbeeldingen die in dit hoofdstuk relevant werden gevonden dat niet meer zijn. Een gebruikersevaluatie zal moeten uitwijzen of de zoekopdrachten voor de afbeeldingen verder verfijnd moeten worden.

4 Implementatie

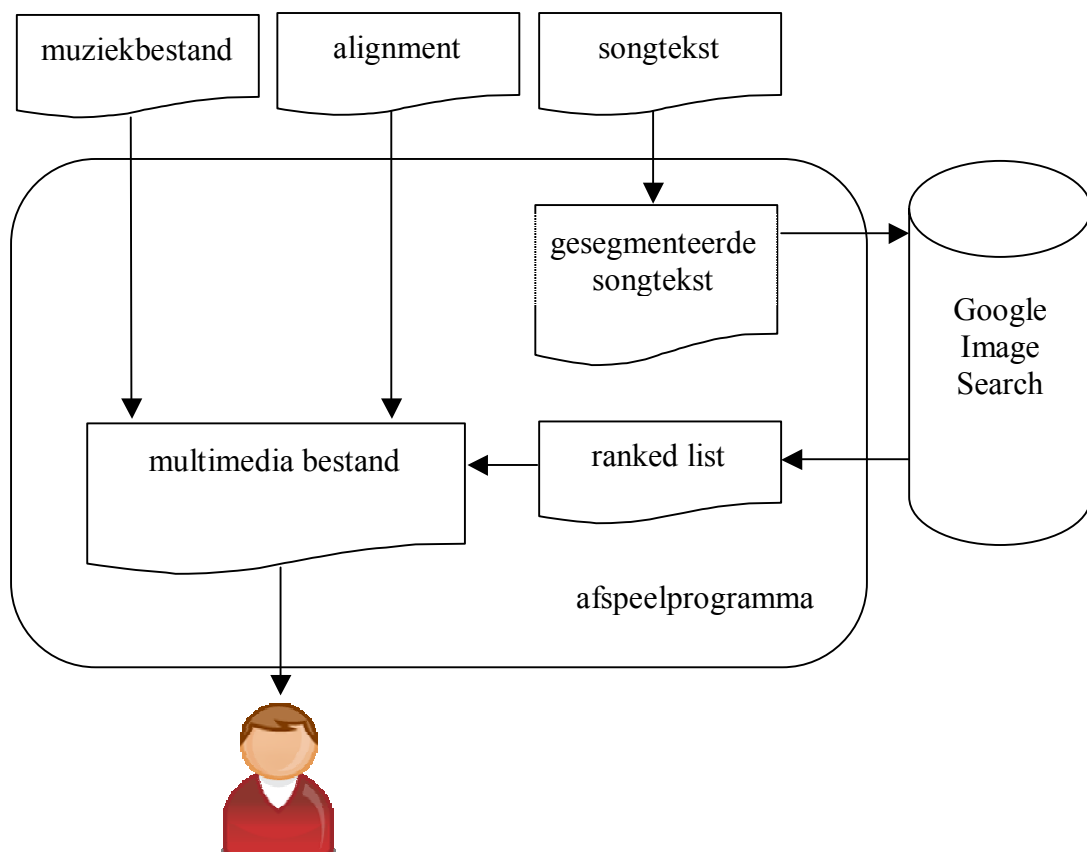
In dit hoofdstuk wordt de implementatie van het afspeelprogramma, de gebruikersevaluatie en de resultaten van de gebruikersevaluatie beschreven.

4.1 Beschrijving van het afspeelprogramma

Het afspeelprogramma voert twee functies na elkaar uit:

- De afbeeldingen worden opgehaald aan de hand van de songtekst
- Het audiobestand wordt afgespeeld en de afbeeldingen worden op het juiste moment getoond aan de gebruiker

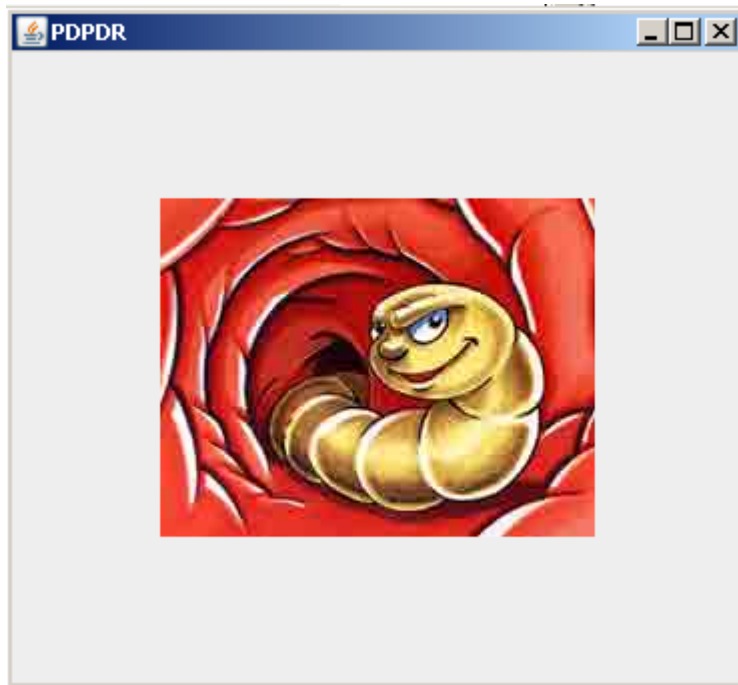
In Figuur 6 wordt schematisch weergegeven wat het afspeelprogramma doet.



Figuur 6: Globale werking van het afspeelprogramma. De songtekst wordt gebruikt om afbeeldingen op te halen uit de Google database. De afbeeldingen, het muziekbestand en de alignment worden als een geheel aan de gebruiker gepresenteerd.

Voordat het audiobestand afgespeeld kan worden moeten de bijbehorende afbeeldingen worden opgehaald. Dit gebeurt zoals beschreven in hoofdstuk 3.2. Bij elke zoekopdracht in de database wordt van de afbeeldingen er willekeurig één gekozen uit de eerste 18 zoekresultaten. Het kan voorkomen dat alle woorden van een tekstregel voorkomen in de stoplijst. Op dat moment kan er dus geen afbeelding bij de tekstregel gevonden worden. In zo'n geval blijft tijdens het afspelen de vorige afbeelding in beeld staan.

Het afspeelprogramma zoekt uit de alignment elk begin van een tekstregel op en het bijbehorende tijdlabel. Tijdens het afspelen wordt zodra er een tijdlabel uit de alignment wordt overschreden een afbeelding in beeld gebracht. Deze afbeelding blijft staan totdat de volgende afbeelding in beeld gebracht moet worden (zie voor een voorbeeld Figuur 7).



Figuur 7: Het afspeelprogramma tijdens het afspelen van het nummer “Elektrostress” van Opgezwolle bij de tekstregel: “dringen ze door naar binnen”

4.2 Gebruikersevaluatie

Om te beoordelen of het toevoegen van de afbeeldingen aan het audiobestand gezien wordt als een verrijking zal er een evaluatie van gebruikers moeten plaatsvinden. In Appendix D wordt een instrument in de vorm van een vragenlijst gegeven. Deze vragenlijst kan gebruikt worden om de gebruikers te vragen naar hun opinie over de verschillende aspecten van het afspeelprogramma.

De vragenlijst heeft twee doelen. De eerste vier vragen proberen meningen van gebruikers te verzamelen over de implementatie van het afspeelprogramma. De respons over die vragen kan door de programmeur gebruikt worden om veranderingen in bijvoorbeeld het zoekalgoritme aan te brengen. De respons over alle vragen kan worden gebruikt om te achterhalen of de toevoeging van afbeeldingen aan muziek gezien wordt als een verrijking.

4.3 Resultaten gebruikersevaluatie

Alhoewel het aantal personen (acht) dat meegedaan heeft aan de gebruikersevaluatie laag is kan het toch een indicatie geven over een aantal aspecten van het afspeelprogramma.

De eerste twee vragen gingen over de relevantie van de afbeeldingen. De gemiddelde score bij de eerste bewering was 2,3 op een schaal van 1 (relevant) tot 6 (niet relevant). Bij de bewering over de relevantie van de afbeeldingen het complete nummer was de gemiddelde score 2,4. De afbeeldingen werden dus relevant gevonden voor de tekstregels.

Over het aantal afbeeldingen was men over het algemeen tevreden. Met de opmerking: “Er werden te weinig afbeeldingen getoond.” was men het niet eens (gemiddelde scores 4,6). Met de complementaire opmerking: “Er werden te veel afbeeldingen getoond.” was men het ook niet eens (gemiddelde score 4,9). Wel werd de opmerking gemaakt dat de afbeeldingen soms te snel weer vervangen werden door een andere afbeelding.

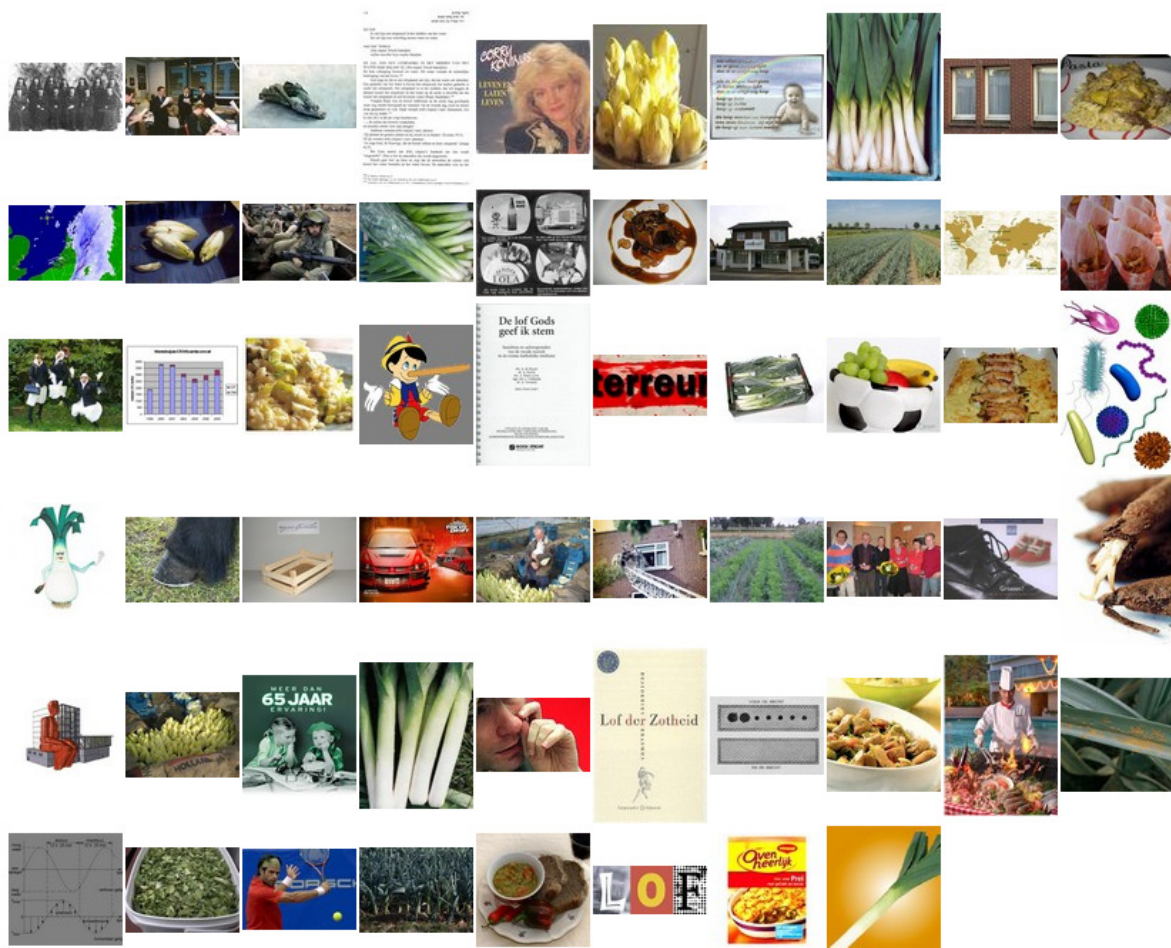
De bewering “De afbeeldingen voegen waarde toe aan het nummer” kreeg een gemiddelde score van 1,9 (men was het eens met die bewering). Ook vond men de combinatie van afbeeldingen en de muziek leuk (gemiddelde score 2,3) en interessant (gemiddelde score 2,6).

Aan het einde van het vragenformulier werden nog twee open vragen gesteld. De eerste open vraag was: “Voor welk soort nummers is deze combinatie van afbeeldingen en muziek geschikt?”. Hierop werd geantwoord dat het afspeelprogramma voornamelijk geschikt voor rustige nummers, met niet te veel tekst en zonder veel beeldspraak. “Je moet de link tussen beeld en geluid meteen kunnen zien.”

De andere open vraag was: “Welke andere toepassingen van deze techniek zijn er volgens u mogelijk?”. Bij deze vraag werd een aantal keren het projecteren van beelden tijdens een voordracht (van bijvoorbeeld een gedicht) of presentatie genoemd. Ook kwam een

aantal keren naar voren om deze techniek te gebruiken om de communicatie met bijvoorbeeld kinderen (plaatjes tijdens verhalen vertellen) of anderstalige mensen te vergemakkelijken. “Als je niet op het Spaanse woord voor lepel kan komen, dan kan je gewoon een plaatje van een lepel laten zien.”.

Een andere interessante opmerking was deze: “Iemand die onbekend met een nummer is, kan via een soort ‘verzamel’ afbeelding (een poster met daarop alle plaatjes die afgebeeld worden tijdens het nummer) snel een indruk krijgen van het nummer. Een duister metal nummer zal naar alle waarschijnlijkheid een andere ‘poster’ hebben dan een vrolijk funk nummer. Hierdoor kun je in een oogwenk beslissen of een nummer iets voor jou is.”. Op deze manier wordt de poster een *visuele representatie* van het complete nummer. Figuur 8 laat zien hoe zo’n poster er uit zou kunnen zien voor het nummer “Knolraap en lof, schorseneren en prei” van Drs. P.



Figuur 8: Visuele representatie van het nummer “Knolraap en lof, schorseneren en prei” van Drs. P.

Om aan een aantal suggesties uit de gebruikersevaluatie te voldoen zou het afspeelprogramma als volgt aangepast kunnen worden:

- In plaats van één afbeelding zou het afspeelprogramma bijvoorbeeld drie afbeeldingen naast elkaar kunnen laten zien zodat er een korte geschiedenis van afbeeldingen in beeld is tijdens het afspelen van een nummer.
- Tijdens het afspelen zou er een poster kunnen worden opgebouwd die dan uiteindelijk een statische visuele representatie van het complete nummer vormt.

Andere mogelijke uitbreidingen zijn:

- In plaats van de tekstregels zou op een triviale manier een heel couplet of refrein als basis kunnen dienen voor de zoekopdrachten. Couplet en refrein worden namelijk meestal in de lay-out van de songtekst gescheiden door een lege regel. Dit zou een door de gebruiker instelbare optie kunnen worden in het afspeelprogramma.
- Met behulp van text clustering zou bepaald kunnen worden welke strofen de coupletten zijn, en welke strofen refreinen zijn. Het afspeelprogramma kan dan voor de refreinen steeds dezelfde afbeeldingen gebruiken. Bij herhaling van het refrein kan er dan ook herhaling van de beelden plaatsvinden tijdens het afspelen van een nummer.
- Het afspeelprogramma zou de meta-informatie die in mediabestanden kan zitten kunnen gebruiken om de zoekopdrachten te tunen. In een zogenaamde mp3-tag kan meta-informatie over het nummer gezet worden zoals het genre, het jaar van de opname, de platenmaatschappij, van welke cd het nummer afkomstig is. Als er bijvoorbeeld een nummer uit 1923 wordt afgespeeld dan zou het afspeelprogramma alleen maar zwart-wit afbeeldingen kunnen tonen.

Als een diepgaande studie nodig is zou het beter zijn om bijvoorbeeld gebruik te maken van de ‘focus group’ methodologie. De focus group methodologie is een kwalitatieve onderzoeksmethode waarbij groepsdiscussies worden georganiseerd om de mening en de ervaringen van mensen te leren kennen over een specifieke set van onderwerpen (Kitzinger, 1994) [20]. Deze methode wordt vaak gebruikt in de marketingindustrie om de mening van

mensen te onderzoeken over een te lanceren nieuw product.

4.4 Conclusies

In dit hoofdstuk is een eerste implementatie beschreven van een afspeelprogramma dat tijdens het afspelen van een audiobestand afbeeldingen kan tonen aan de gebruiker. De afbeeldingen worden getoond aan de hand van de tijdlabele verkregen uit de alignment zoals beschreven in hoofdstuk 2. De afbeeldingen worden uit een online database gehaald zoals beschreven in hoofdstuk 3.

Er is een korte gebruikersevaluatie uitgevoerd waaruit blijkt dat men deze combinatie van afbeeldingen en geluid leuk en interessant vindt. Ook vindt men over het algemeen de afbeeldingen bij het nummer passen. Verder heeft de gebruikersevaluatie bruikbare suggesties opgeleverd voor uitbreiding van het programma en eventueel verder onderzoek.

5 Conclusies en aanbevelingen

In dit onderzoek is een spraakherkenner gebruikt om tijdlabls voor songteksten te genereren door middel van forced alignment. Voor vijftien muziknummers is de kwaliteit van deze tijdlabls geëvalueerd. Er is een afspeelprogramma gemaakt dat aan de hand van de tekstregels van de songtekst worden in een database relevante afbeeldingen opzoekt. Deze afbeeldingen worden op het juiste moment (aan de hand van de tijdlabls) aan de gebruiker getoond. Tot slot is het afspeelprogramma door gebruikers geëvalueerd.

5.1 *Conclusies*

Het is handig om de vragen die aan het begin van dit verslag in sectie 1.1 werden gesteld er bij te halen om te kijken of ze beantwoord zijn:

- Kan een (gesegmenteerde) songtekst automatisch voorzien worden van tijdinformatie?

In dit onderzoek is een evaluatiemethode uitgewerkt om de kwaliteit van een alignment te beoordelen. Na parametrisering van de spraakherkenner zijn de alignment-scores voor een vijftiental nummers bepaald. Negen van de vijftien nummers hadden een goede alignment. Voor nummers met een slechte alignment blijkt het te helpen als het nummer wordt opgesplitst in delen en dat daarna die delen apart worden opgelijnd. Dat leverde nog eens vier nummers op met een goede alignment. De automatisch gegeneerde tijdlabls zijn dus in veel gevallen geschikt voor gebruik in het afspeelprogramma om op de juiste momenten de afbeeldingen op het scherm te tonen.

- Op welke manier kan de koppeling tussen afbeeldingen en de songtekst gemaakt worden?

Dit is mogelijk door elke tekstregel van een songtekst te gebruiken als basis voor een zoekopdracht in een afbeeldingendatabase.

- Is het mogelijk om passende beeldinformatie toe te voegen aan een muziekstuk?

De tekstregels die als zoekopdracht worden gebruikt voor de afbeeldingendatabase moeten dan eerst gefilterd worden met behulp van een stoplijst. Ook moeten er een aantal opties meegegeven worden aan de query om de resultaten uit de database relevanter te maken.

- Wordt deze toevoeging inderdaad gezien als een verrijking?

Om deze vraag te kunnen beantwoorden is er een korte gebruikersevaluatie uitgevoerd waaruit bleek dat men deze combinatie van afbeeldingen en geluid leuk en interessant vond. Ook vond men over het algemeen de afbeeldingen bij het nummer passen.

5.2 Aanbevelingen

Aanbevelingen voor mogelijke aanpassingen van de gebruikte zoekopdrachten zijn:

- In plaats van de tekstregels zou een heel couplet of refrein als basis kunnen dienen voor de zoekopdrachten.
- Het afspeelprogramma zou de meta-informatie die in mediabestanden kan zitten kunnen gebruiken om de zoekopdrachten te tunen.

Aanbevelingen voor het uitbreiden van het afspeelprogramma zijn:

- Het afspeelprogramma zou uitgebreid kunnen worden tot een karaoke-machine. Hiervoor moet de automatische alignment tot op woord-niveau goed werken.
- Met behulp van text clustering zou bepaald kunnen worden welke strofen de coupletten zijn, en welke strofen refreinen zijn. Het afspeelprogramma kan dan voor de refreinen steeds dezelfde afbeeldingen gebruiken.
- In plaats van één afbeelding zou het afspeelprogramma bijvoorbeeld drie afbeeldingen naast elkaar kunnen laten zien zodat er een korte geschiedenis van afbeeldingen in beeld is tijdens het afspelen van een nummer.

Aanbevelingen voor andere mogelijke toepassingen van dezelfde techniek:

- Eén van de suggesties was om een poster van de afbeeldingen te maken zodat er een statische visuele representatie van het complete nummer wordt gevormd. Wellicht kan hier meer mee gedaan worden.
- Een andere suggestie was om het systeem te gebruiken voor communicatie. Het is mogelijk om een bestaande ‘instant messenger’ aan te passen en de technieken in dit onderzoek toe te passen op de berichten die gebruikers naar elkaar toe sturen.

6 Referenties

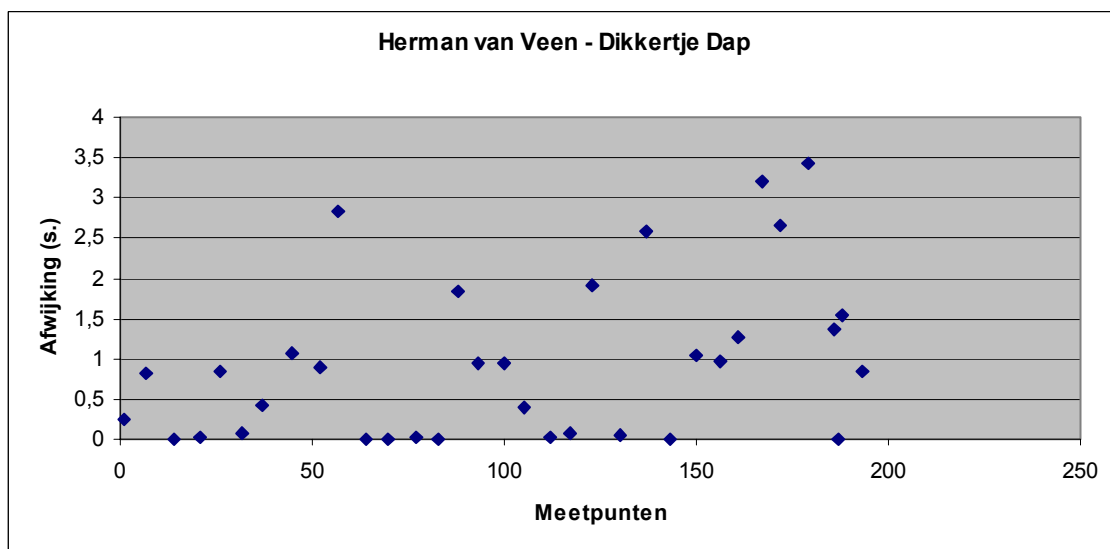
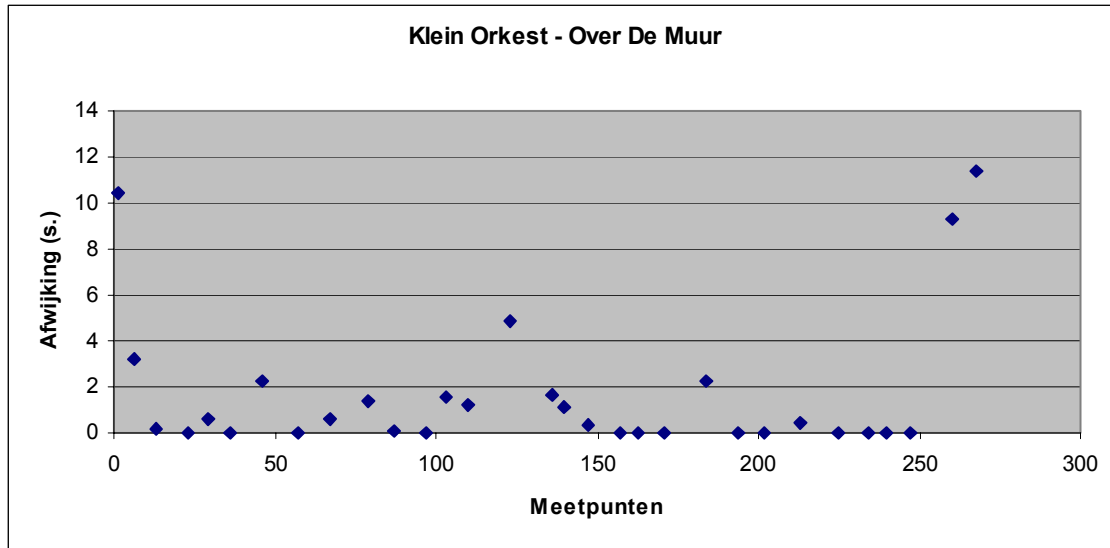
- [1] *Nieuwste FlatTV's van Philips intensiveren kijkervaring als nooit tevoren*, Philips website, laatst bezocht op 18-08-2007, <http://www.philips.nl/about/news/section-13669/article-14437.html>
- [2] Franssen, M. [1991], '*The Ocular Harpsichord of Louis-Bertrand Castel: The Science and Easthetics of an Eighteenth-Century Cause Célèbre*', Tractrix. Yearbook for the History of Science, Medicine, Technology and Mathematics 3 (1991), 15-77.
- [3] Bishop, B. [1893], *A Souvenir of the Color Organ, with some Suggestions in Regard to the Soul of the Rainbow and the Harmony of Light*, New Russia, New York., beschikbaar op rhythmiclight website, laatst bezocht op 26-07-2007, <http://www.rhythmiclight.com/archives/index.html>
- [4] *A Brief History of Synesthesia in the Arts*, comcast website, laatst bezocht op 26-07-2007, <http://home.comcast.net/~sean.day/art-history.htm>
- [5] Løvstrøm, R.E., *Artists on line: Thomas Wilfred and his Clavilux*, gis.net website, laatst bezocht op 26-07-2007, <http://www.gis.net/~scatt/clavilux/clavilux.html>
- [6] Popular Mechanics, [april 1924], *Music Is Turned Into Glowing Color: Soundless Symphonies from Keys of "Organ" Projected on Screen Are Hailed as Birth of a New Art.*
- [7] *Lumia – Thomas Wilfred*, clavilux website, laatst bezocht op 26-07-2007, <http://www.clavilux.org/>
- [8] Rees, A.L. [1999], *A History of Experimental Film and Video*, British Film Institute.
- [9] Decker-Phillips, E. [1998], *Paik Video*, Barrytown Ltd.
- [10] Ye Wang, Min-Yen Kan, Tin Lay New, Arun Shenoy, and Jun Yin [2004], *LyricAlly*:

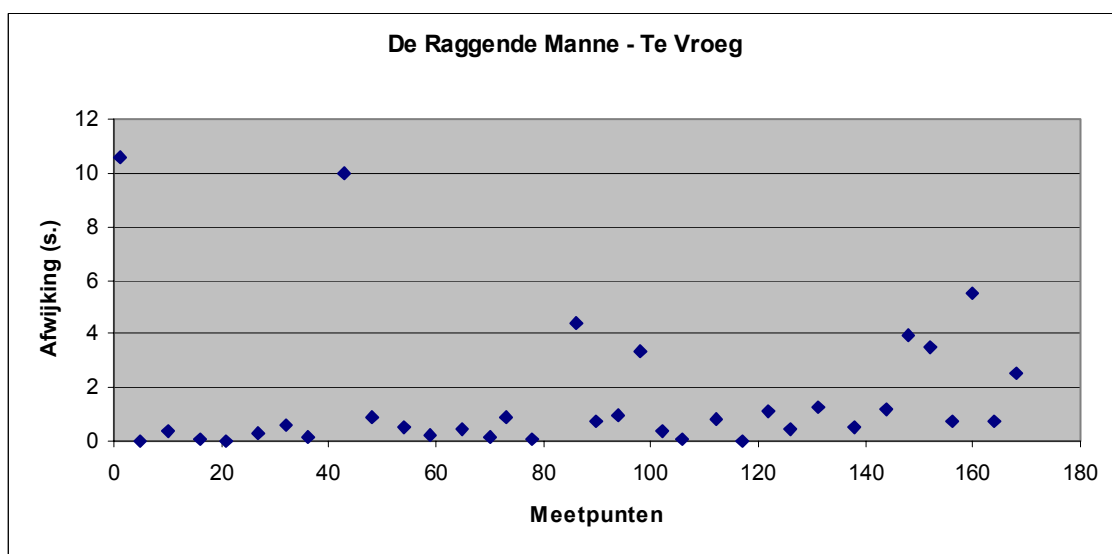
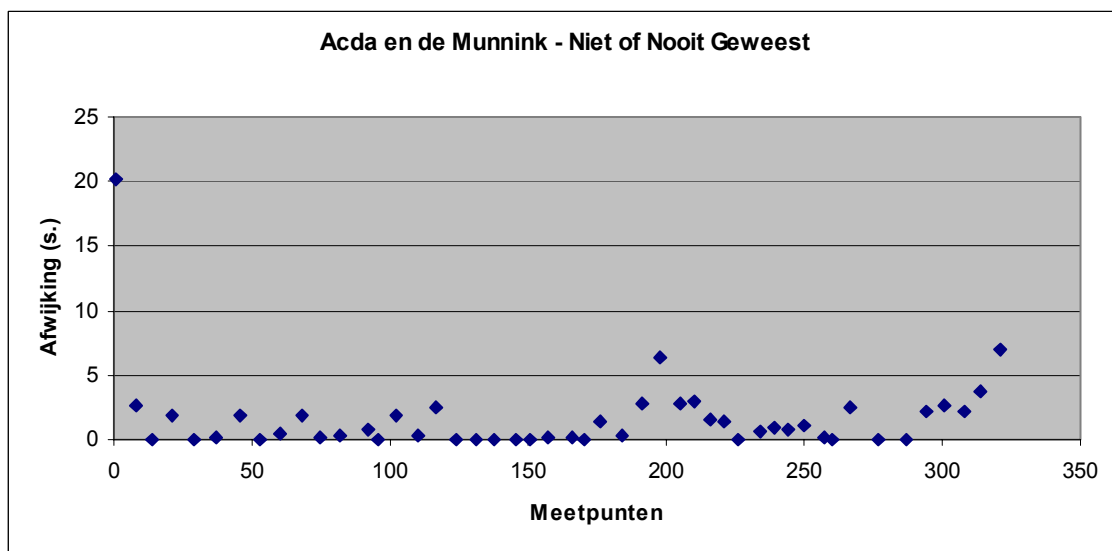
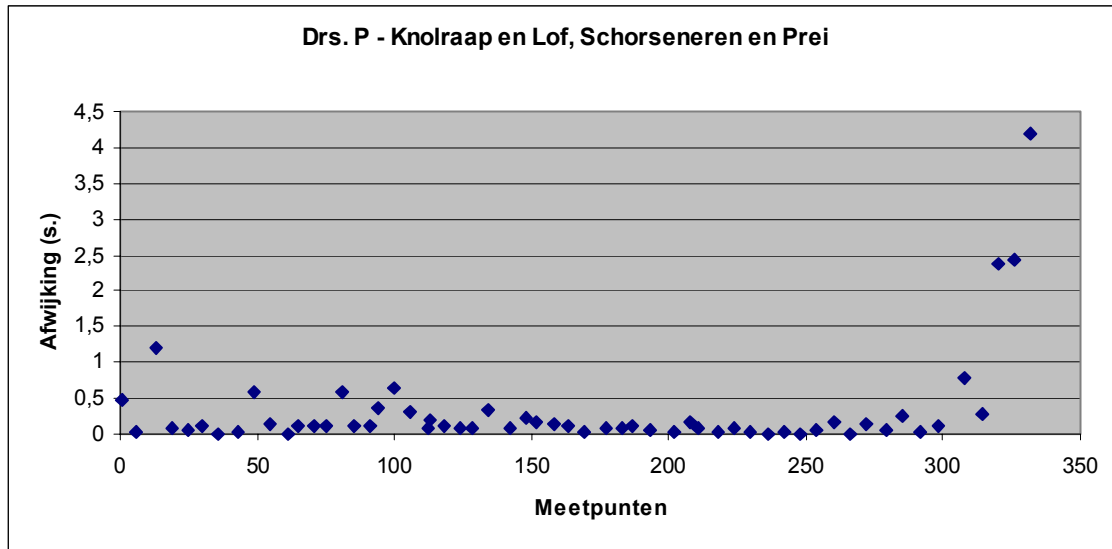
- Automatic Synchronization of Acoustic Musical Signals and Textual Lyrics*, In Proc. Of ACM Multimedia '04, New York, NY, USA
- [11] Pellom, B. [2001], *SONIC: The University of Colorado Continuous Speech Recognizer*, Boulder, Colorado, US, University of Colorado.
- [12] Nair, G. [2007], *The craft of singing*, San Diego, CA: Plural Publishing.
- [13] *Transcriber: a tool for segmenting labeling and transcribing speech*, sourceforge website, laatst bezocht op 26-07-2007, <http://trans.sourceforge.net/en/presentation.php>
- [14] Knees, P., M. Schedl, and G. Widmer [2005], *Multiple lyrics alignment: Automatic retrieval of song lyrics*, Proceedings 6th International Conference on Music Information Retrieval, ISMIR, London, UK, 564–569.
- [15] Korst, J., and G. Geleijnse [2006], *Efficient Lyrics Retrieval and Alignment*.
- [16] L. Chen, Y. Liu, M. Harper, E Maia, and S. McRoy [2004], *Evaluating Factors Impacting the Accuracy of Forced Alignments in a Multimodal Corpus*.
- [17] *Flickr - Share your photos Watch the world*, Flickr website, laatst bezocht op 26-07-2007, <http://www.flickr.com/>
- [18] *Zoomr – Eindeloos Foto's Delen*, Zoomr website, laatst bezocht op 26-07-2007, <http://www.zoomr.com/>
- [19] *Google Afbeeldingen Zoekmachine*, Google website, laatst bezocht op 26-07-2007, <http://images.google.com/>
- [20] Kitinger, J. [1994], 'The methodology of Focus Groups: the importance of interaction between research participants', *Sociology of Health and Illness*, v.16, No1, pp: 103-121.

APPENDIX

A Grafieken Categorie I

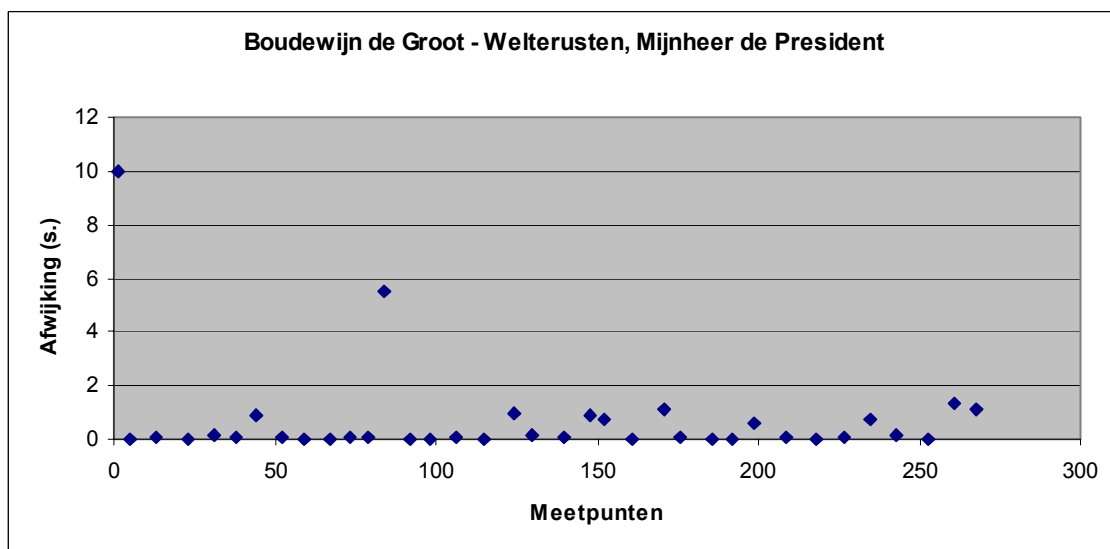
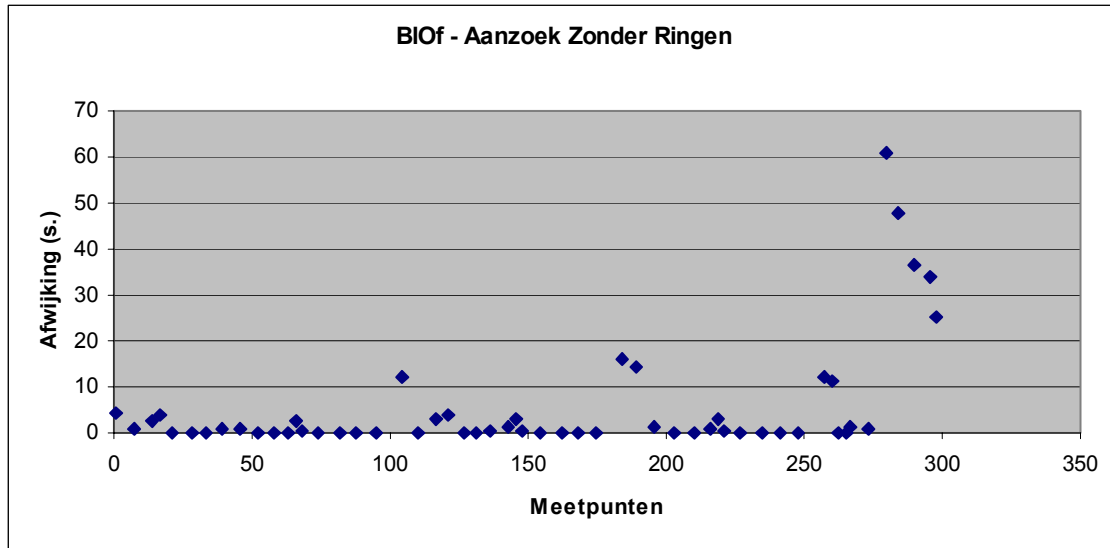
In deze appendix wordt voor elk nummer uit de eerste categorie (sectie 2.6.1) de afwijking tussen de door de alignment-tool opgeleverde starttijden van een tekstregel en de ground truth in een grafiek weergegeven.

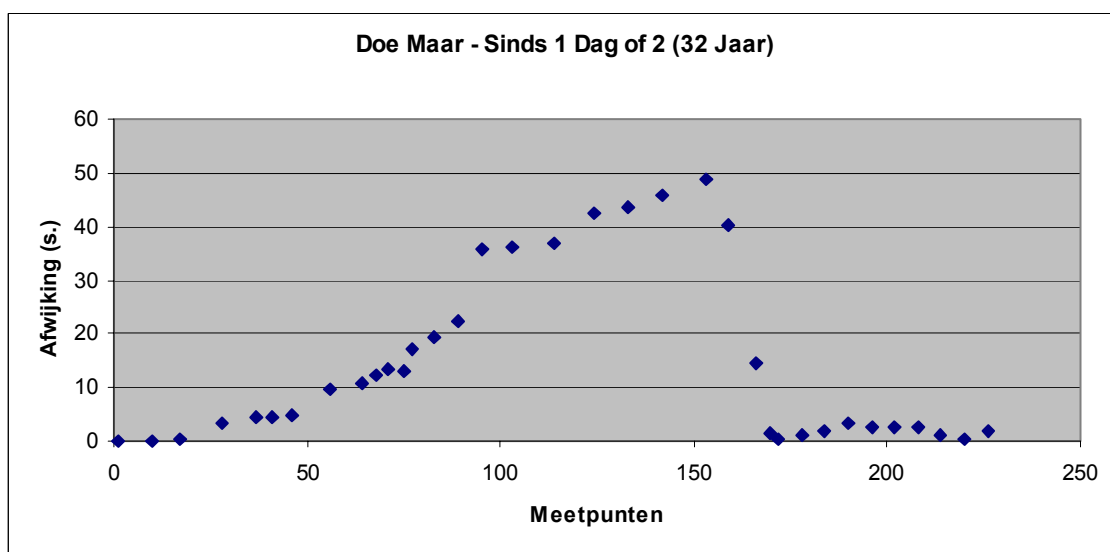
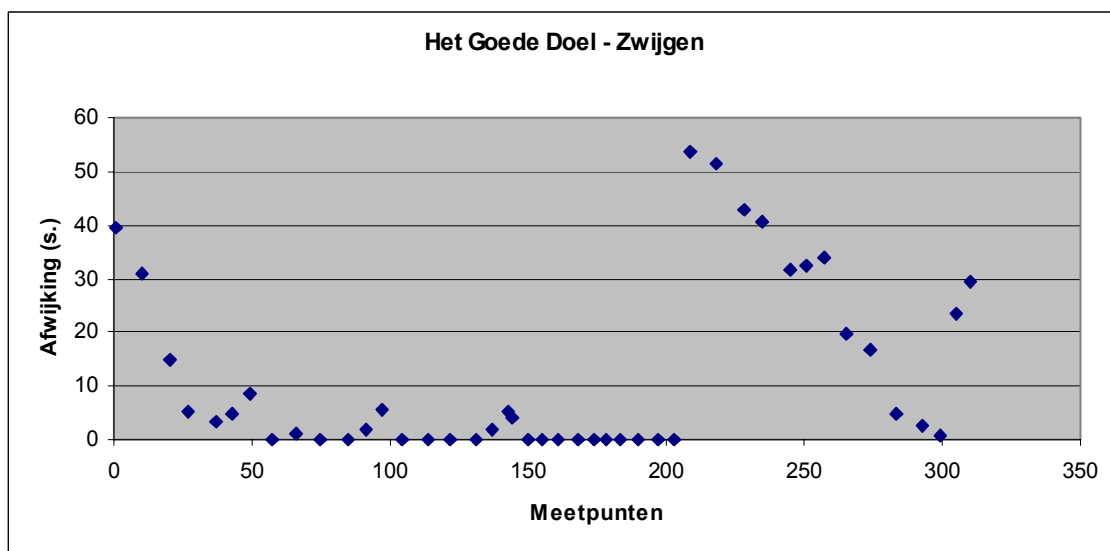
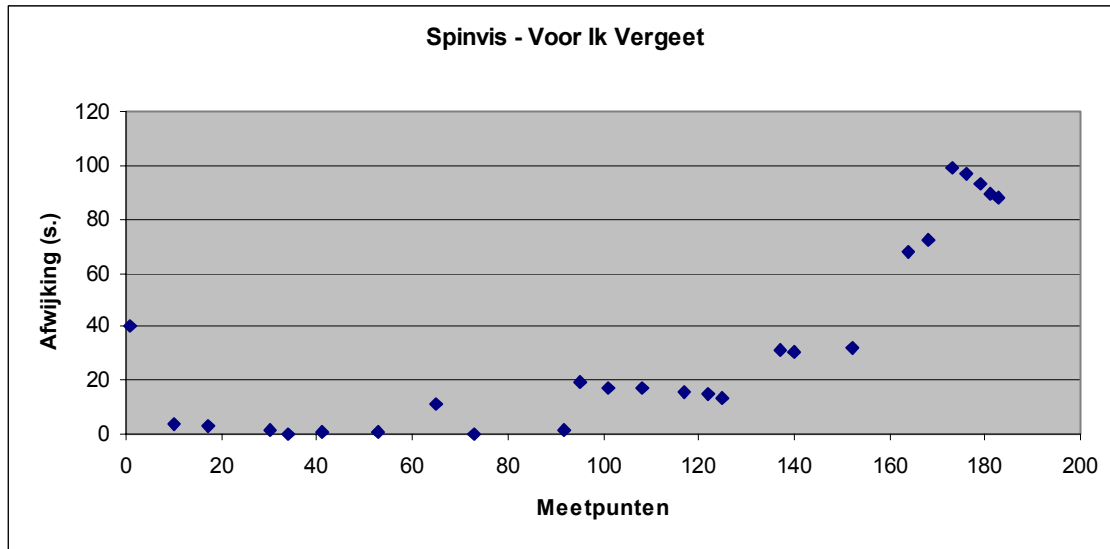




B Grafieken Categorie II

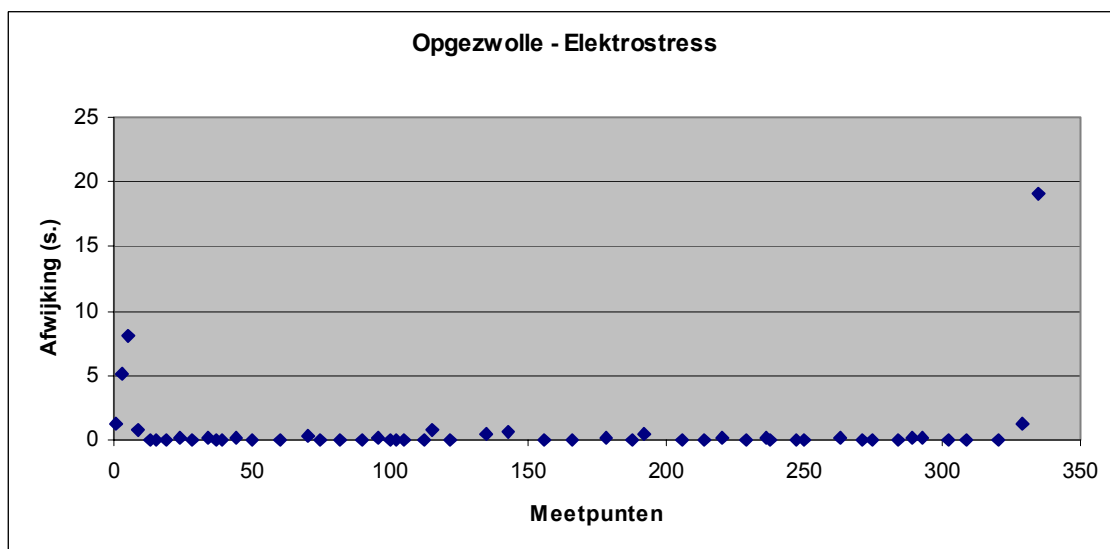
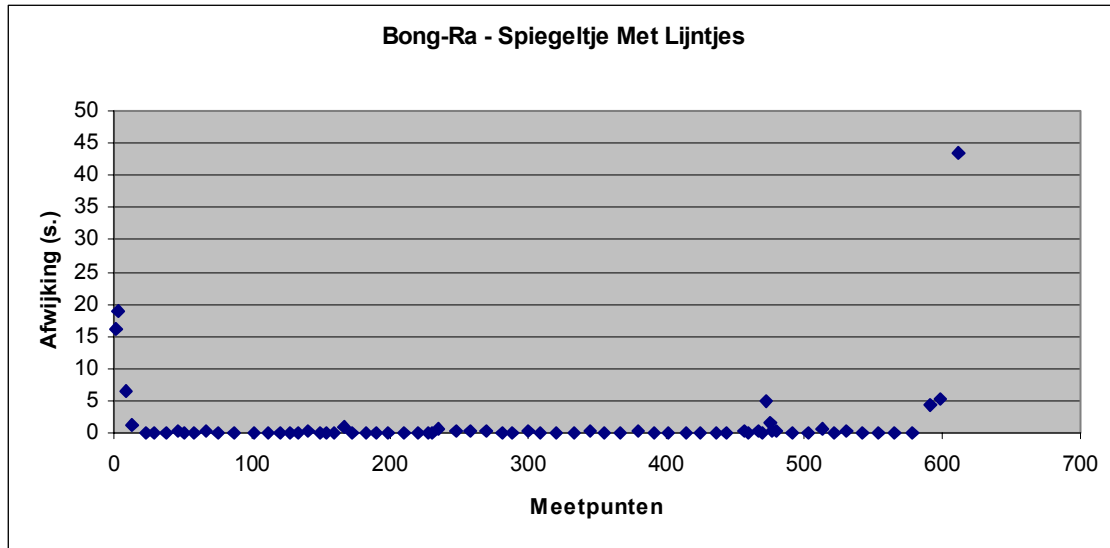
In deze appendix wordt voor elk nummer uit de tweede categorie (sectie 2.6.2) de afwijking tussen de door de alignment-tool opgeleverde starttijden van een tekstregel en de ground truth in een grafiek weergegeven.

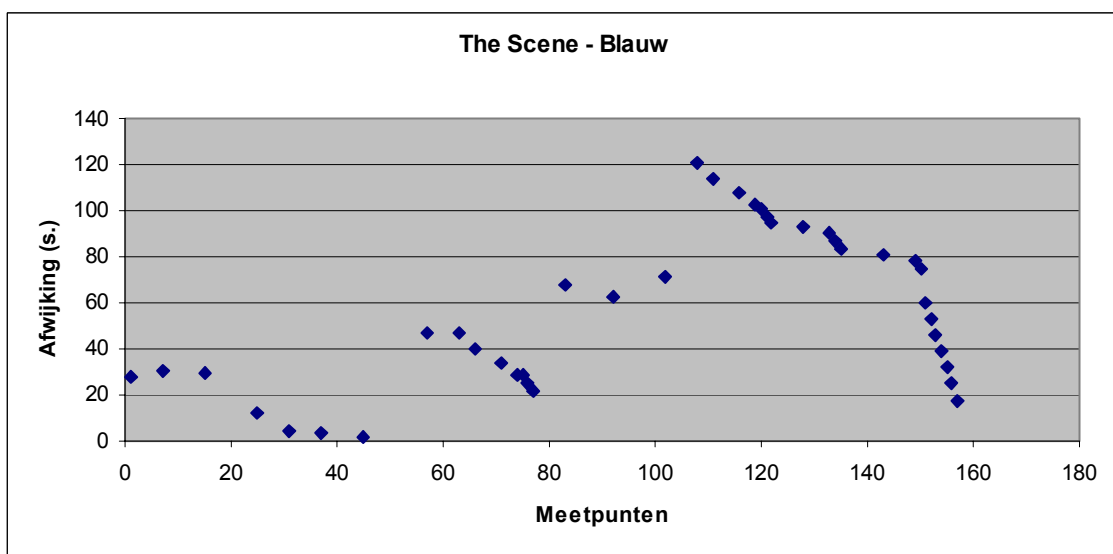
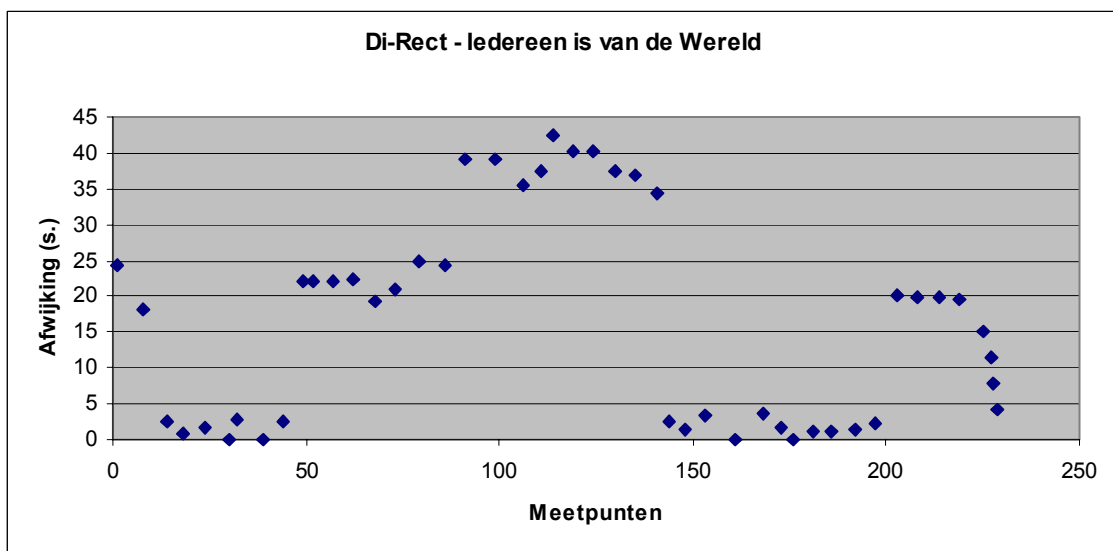
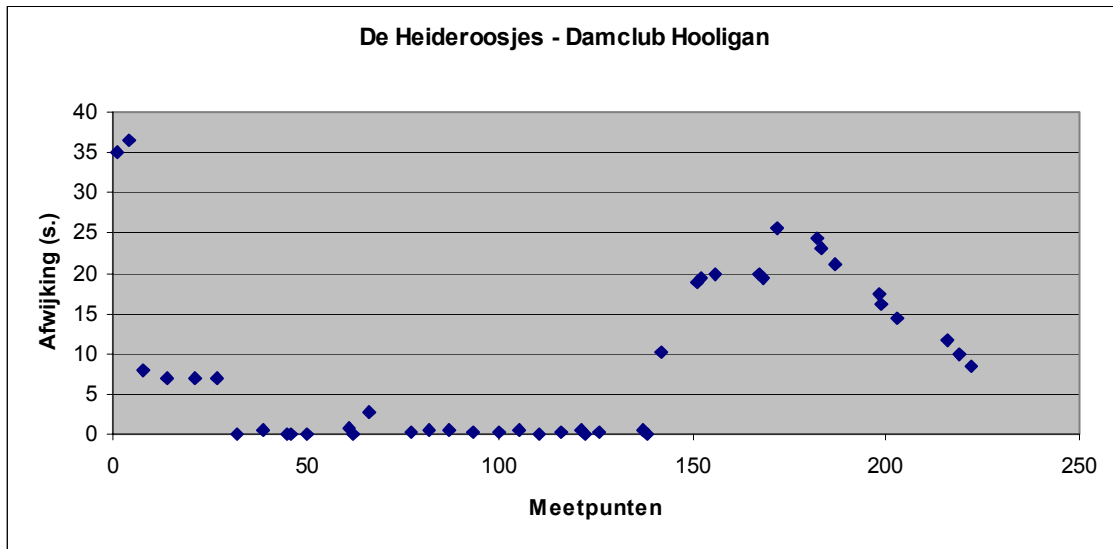




C Grafieken Categorie III

In deze appendix wordt voor elk nummer uit de derde categorie (sectie 2.6.3) de afwijking tussen de door de alignment-tool opgeleverde starttijden van een tekstregel en de ground truth in een grafiek weergegeven.





D Formulier gebruikersevaluatie

Datum:

Deelnemersnummer:

Geachte deelnemer,

Onderstaand formulier maakt onderdeel uit van een onderzoek naar de mogelijkheden om aan muziek beelden toe te voegen op basis van de songtekst. We willen u vragen om deel te nemen aan het onderzoek door onderstaande vragen te beantwoorden. De resultaten zullen anoniem worden behandeld dus voel u vrij om uw echte mening te geven.

Instructie:

U krijgt straks twee keer een nummer te horen. U krijgt het nummer de eerste keer zonder beelden te horen om u bekend te laten worden met het nummer. De tweede keer worden er tijdens het afspelen van het nummer afbeeldingen getoond.

We vragen u om daarna uw opinie over een aantal beweringen te geven. U kunt uw opinie geven door een X te zetten in dat gedeelte van de schaal dat uw mening over de bewering het beste weergeeft. Aan het eind volgen er nog twee open vragen.

Voorbeeld:

Bijvoorbeeld, in de volgende opmerking.

Ik houd van ijs met chocoladesmaak

--	--	--	--	--	--

Helemaal mee eens

Helemaal niet mee eens

Als u erg van ijs met chocoladesmaak houdt, dan zet u een X aan de linkerkant zoals hieronder:

X					
---	--	--	--	--	--

Helemaal mee eens

Helemaal niet mee eens

Echter, als u totaal niet van ijs met chocoladesmaak houdt, dan zet u een X aan de rechterkant van de schaal, zoals hieronder:

					X
--	--	--	--	--	---

Helemaal mee eens Helemaal niet mee eens

U kunt ook een X op een andere plek van de schaal zetten al naar gelang de beweringen met uw opinie overeenkomen.

Als u nog vragen heeft over hoe u de questionnaire moet beantwoorden, twijfel niet om ons te vragen.

Zet een X op de volgende schalen al naar gelang uw mening over de getoonde beweringen.

In het zojuist afgespeelde nummer:

1. De afbeeldingen waren relevant voor *de tekstregels*

--	--	--	--	--	--

Helemaal mee eens Helemaal niet mee eens

2. De afbeeldingen waren relevant voor *het complete nummer*

--	--	--	--	--	--

Helemaal mee eens Helemaal niet mee eens

3. Er werden te weinig afbeeldingen getoond

--	--	--	--	--	--

Helemaal mee eens Helemaal niet mee eens

4. Er werden te veel afbeeldingen getoond

--	--	--	--	--	--

Helemaal mee eens

Helemaal niet mee eens

5. De afbeeldingen voegen waarde toe aan het nummer

--	--	--	--	--	--

Helemaal mee eens

Helemaal niet mee eens

6. Wat was uw algemene indruk over de combinatie van afbeeldingen en de muziek?

--	--	--	--	--	--

Erg leuk

Niet leuk

--	--	--	--	--	--

Erg interessant

Erg saai

Beantwoord de volgende open vragen.

7. Voor welk soort nummers is deze combinatie van afbeeldingen en muziek geschikt?

8. Welke andere toepassingen van deze techniek zijn er volgens u mogelijk?

HARTELIJK DANK VOOR UW DEELNAME!