
Samenvatting

De basis van dit onderzoek is gelegd in de periode tussen 1994 en 1999 toen ik een onderwijsaanstelling had bij de studierichting Technische Cognitiewetenschap. Doel van het onderzoek was het vinden van spraakherkenningstoepassingen van de rijke temporele informatie in een model van het menselijk binnenoor of cochlea dat ontwikkeld was binnen de groep van prof. Duifhuis. Op het moment dat de eerste resultaten suggereerden dat op basis van dit onderzoek belangrijke toepassingsmogelijkheden mogelijk waren in, onder andere, de spraaktechnologie, is besloten om te wachten met publikaties en één en ander verder te ontwikkelen ten behoeve van octrooiaanvragen. Eind 1999 is, op initiatief en met financiële steun van de RuG Houdstermaatschappij besloten om het onderzoek voort te zetten in bedrijfsverband. Hiertoe is het bedrijf *Human Quality (HuQ) Speech Technologies BV* opgericht.

Continuity preserving Signal Processing (CPSP) is een geheel nieuwe signaalanalysemethodologie die gebaseerd is op het onderzoek dat beschreven is in dit proefschrift. Daar CPSP is gebaseerd op algemeen geldende basisaannames is het, in tegenstelling tot andere signaalanalysemethoden, geschikt voor de analyse van onbekende (d.w.z. niet herkende) complexe geluidssignalen. Met behulp van CPSP zijn topresultaten mogelijk op de Aurora test: een internationale *benchmark* test op het gebied van ruisrobuuste spraakherkenning. Dit onderzoek heeft zich ontwikkeld als enerzijds de aanzet voor een theorie over menselijke (spraak)geluidsanalyse en anderzijds een nieuwe benadering van spraaktechnologie. Mede om deze redenen is ervoor gekozen

om het proefschrift te schrijven als een *Tutorial* die zich in de loop van de tijd verder zal ontwikkelen.

Hoofdstuk 1 - Recognizing Arbitrary Sounds

Dit hoofdstuk richt zich op de vraag wat de eisen zijn die aan een herkenningssysteem (natuurlijk of artificieel) gesteld moeten worden om te garanderen dat de systemen in zoveel mogelijk omstandigheden betrouwbaar kunnen functioneren. Allereerst wordt ingegaan op het feit dat de huidige generatie automatische spraakherkenners (ASR-systemen) niet de veelzijdigheid en betrouwbaarheid vertonen die het menselijk auditief systeem kenmerkt. ASR-systemen falen, of zijn zeer beperkt in hun mogelijkheden, in situaties waar de menselijke gebruiker geen enkel probleem heeft. Dit leidt tot een sterk verschil tussen de verwachtingen van de gebruiker en de mogelijkheden van ASR-systemen; en uiteindelijk tot teleurgestelde en ontevreden gebruikers. De (markt)vraag om de ruisrobuustheid (hier gedefinieerd als de betrouwbare werking in situaties waarin het menselijk systeem probleemloos functioneert) van spraakherkenningssystemen te vergroten motiveert zeer veel onderzoek.

Dit onderzoek is gebaseerd op twee aannames. Ten eerste dat ons auditief systeem is geoptimaliseerd om in zoveel mogelijk situaties adequaat te functioneren. Ten tweede dat de belangrijkste linguïstische informatie gerepresenteerd wordt door de meest ruisrobuuste signaalcomponenten. Het tweede uitgangspunt maakt het mogelijk om bij het analyseren van geluidssignalen (die al of niet spraak kunnen bevatten), niet in eerste instantie te zoeken naar belangrijke linguïstische componenten, maar juist naar *ruisrobuuste* componenten. Een verdere analyse van gevonden ruisrobuuste (en dus betrouwbare) componenten kan leiden tot een valide linguïstische interpretatie. Volgende hoofdstukken richten zich op de analyse en detectie van ruisrobuuste signaalcomponenten.

De huidige generatie ASR-systemen, op basis van *Hidden Markov Modellen* (HMM) is oorspronkelijk ontwikkeld voor het herkennen van “schone” spraak, d.w.z. signalen met uitsluitend (doel)spraak. Omdat deze bijzondere situatie meestal niet gegarandeerd kan worden zijn er - met beperkt succes - technieken ontwikkeld om de systemen toch te laten functioneren in ruizige situaties. Een HMM-gebaseerd ASR-systeem gaat ervan uit dat signaalverwerking en patroonherkenning gescheiden processen zijn, die na elkaar plaatsvinden. In de meeste ruizige situaties echter, is het uitsluitend mogelijk om de juiste signaalcomponenten te detecteren en te combineren indien bekend is wat er gezocht moet worden. Helaas kan het systeem dit nog niet we-

ten, want er is nog geen herkenningresultaat. Deze paradoxale situatie, die ik *de spraak-in-ruis-paradox* noem, is een direct gevolg van de scheiding van signaalverwerking en patroonherkenning, en is, mijns inziens, de reden dat de huidige generatie ASR-systemen de menselijke norm niet kunnen benaderen.

Een echt ruisrobuust herkenningssysteem (zoals de mens) vermijdt de paradoxale situatie door het geluidssignaal eerst te ontleden in (ruisrobuuste) signaalcomponenten en die vervolgens te combineren in representaties die voldoen aan de karakteristieke eigenschappen van de klasse waar ze een instantie van zijn. Met andere woorden: bij het herkennen van spraak mogen signaalcomponenten uitsluitend gecombineerd worden indien ze consistent zijn met de karakteristieke eigenschappen van spraakklanken en sprekers. Het proefschrift richt zich in eerste instantie op het ontleden van geluiden in spraakcomponenten en het combineren van de componenten op basis van eenvoudige fysische eigenschappen. Het formuleren van de juiste spraak en spreker specifieke eigenschappen is nog onderwerp van onderzoek.

Het hoofdstuk wordt voortgezet met een pleidooi voor het behoud van continuïteit bij de analyse van onbekende signalen. Standaard spraaksignaalanalyse technieken (meestal op basis van Fourier Analyse) delen het signaal op in *frames* die elk de informatie over een beperkt tijdsinterval vertegenwoordigen. Dit kan gerechtvaardigd worden op grond van de quasistationariteitsaannname die zegt dat, indien een signaal langzaam genoeg verandert, het bemonsterd kan worden met een frequentie die afhangt van de dynamiek van het signaal. De dynamische mogelijkheden van spraaksignalen worden beperkt door de traagheid van het stemkanaal. Dit rechtvaardigt een "bemonster periode" van 5 tot 10 ms voor het beschrijven van veranderingen in het stemkanaal van een enkele spreker.

Van niet herkende geluidssignalen is het onbekend of, en zo ja op welke wijze, quasistationariteit toegepast kan worden. Dit leidt tot de conclusie dat de toepassing van quasistationariteit uitgesteld dient te worden tot het moment dat de veilige toepassing ervan gegarandeerd is. Tot die tijd moet de continuïteit in het signaal behouden blijven. Dit is precies wat lijkt te gebeuren in de eerste stap van het menselijk auditief systeem. Sterke signaalcomponenten zijn in staat om gedurende hun hele bestaan hun periodiciteit op te leggen aan het basilair membraan (de plaats waar geluidstrillingen omgezet worden in neurale informatie). Dit betekent dat deze geluidscomponenten het basilair membraan lokaal domineren en daarmee gedurende hun bestaan een bron van betrouwbare informatie blijven vormen.

Hoofdstuk 1 wordt afgesloten met een formulering van de taak van de rest van het proefschrift. Het gaat om:

1. het vaststellen van de delen van een schoon signaal die, in ruizige omstandigheden, waarschijnlijk een hoge signaal-ruis-verhouding hebben;
2. het selecteren van deze ruisrobuuste delen uit een verruiste versie van het signaal; en
3. het vaststellen of de energie- en frequentieontwikkeling van de geselecteerde delen van het verruiste signaal in overeenstemming is met die van vergelijkbare delen van het schone signaal.

De hypothesen geformuleerd in hoofdstuk 1 vormen de basis voor de rest van het proefschrift.

Hoofdstuk 2: Continuity Preserving Signal Processing

De taak van dit hoofdstuk is tweeledig. Het geeft een overzicht van de basale representaties van *Continuity Preserving Signal Processing* (CPSP) en het beschrijft de eerste herkenningsexperimenten op basis van CPSP technieken in combinatie met een HMM herkenner. Allereerst worden CPSP technieken en representaties en hun onderlinge relaties weergegeven in een overzicht, waarin voor periodieke signalen een speciaal traject bestaat. Op basis van dit traject wordt in het hoofdstuk uiteengezet hoe het mogelijk is om, in ruizige omstandigheden, veel van de energie die afkomstig is van een periodiek signaal te selecteren en om te zetten in een representatie die geschikt is als input voor een herkenningssysteem.

De beschrijving van CPSP wordt begonnen met een korte beschrijving van de essentiële (continuïteit behoudende) eigenschappen van het lineaire basilair membraanmodel dat wordt gebruikt. De output van het model kan met behulp van een lekkend integratieproces worden omgezet in een *cochleogram*: een representatie van energie als continue functie van tijd en plaats (waarbij plaats correspondeert met frequentie) die vergelijkbare informatie vertegenwoordigt als een standaard spraakspectrogram. De zogenaamde *Tuned Auto-correlation* (TAC) is een variant van het cochleogram waarbij in de definitie de basisperiode als signaaleigenschap is opgenomen. Wanneer de periodecontour correct gekozen is fungeert de TAC als een filter dat alle bewaard gebleven informatie van de met de periodecontour corresponderende bronnen selecteert. De TAC is een zeer efficiënte (doch niet feilloze) manier om in een ruizig signaal vast te stellen welke delen van het basilair membraan worden gedomineerd door periodieke delen van het doelsignaal.

Zowel het cochleogram als de TAC zijn bijzondere deelverzamelingen van de *Time Normalized Correllogram* (TNC), een generalisatie van de autocorrelatie die continuïteit garandeert in tijd, plaats en periodiciteit. De temporele normalisatie houdt in dat het tot expressie komen van de start van signaalcomponenten pas begint wanneer er precies één periode beschikbaar is. De eigenschappen van de TNC vormen de kern van CPSP. Daar de TNC een zeer redundante representatie is, is het van belang om de TNC op een efficiënte wijze te doorzoeken. Het cochleogram, de TAC en de in hoofdstuk 4 te introduceren *Characteristic Period Correlation* (CPC) zijn samen met de voortschrijdende autocorrelatie (*running autocorrelation*) langs een “ridge” de meest informatieve deelverzamelingen van de TNC.

Ridges zijn reeksen van pieken in het cochleogram die ontstaan wanneer een periodieke signaalcomponent (meestal een harmonische van een quasiperiodiek signaal) tijdelijk een deel van het basilaire membraan domineert. *Ridges* zijn cochleogramposities waar quasiperiodiciteit bij ruizige signalen veilig toepast kan worden. Door onder een *ridge* een zich in de tijd ontwikkelende autocorrelatie te berekenen is het mogelijk om de locale instantane frequentieontwikkeling van de signaalcomponent zeer nauwkeurig te volgen. Dit leidt tot *Locale Instantane Frequentie* (LIF) *contouren* die aan individuele signaalcomponenten gekoppeld zijn. Van de LIF *contouren* wordt vastgesteld dat ze in een ruizige situatie zeer robuust en betrouwbaar zijn.

Na deze introductie in basale CPSP representaties wordt een testsysteem beschreven dat ontwikkeld is om aan te tonen dat de TAC een geschikte signaalselectiemethode is om in ruizige omstandigheden informatie van een doelsignaal te selecteren en te herkennen met een standaard ASR-systeem. Hiervoor zijn periodecontouren nodig die bepaald kunnen worden door LIF *contouren* te combineren. Hoewel dit leidt tot valide periodecontouren kan op basis van de beschikbare informatie niet met zekerheid vastgesteld worden dat de gevonden contour van het doelsignaal zijn. (Ad ultimo een consequentie van de signaal-in-ruis-paradox.)

De basilaire membraam oscillaties die corresponderen met cochleogram oppervlakken die met behulp van de TAC kunnen worden geselecteerd, kunnen worden gebruikt om via een invers filteringsproces een geluidssignaal te resynthetiseren. In het geresynthetiseerde signaal is een groot deel van de ruis energie niet meer aanwezig. Dit toont aan dat het inderdaad mogelijk is om op basis van CPSP technieken informatie van één bron op efficiënte wijze te selecteren uit een mengsel van geluidssignalen.

In het laatste deel van het hoofdstuk wordt toegewerkt naar een efficiënte parameterisatie van TAC selecties, waarna de herkenningresultaten (op basis van correcte, in schone omstandigheden vastgestelde periodecontouren) aantonen dat de TAC selecties inderdaad leiden tot een zeer ruisrobuust herkenningresultaat.

Hoofdstuk 3: The Basilar Membrane Response

De volgende hoofdstukken gaan in op details en uitbreidingen van de technieken die in hoofdstuk 2 zijn geïntroduceerd. Hoofdstuk 3 gaat allereerst in de op de implementatie van het cochleamodel (Duijfhuis, et.al. 1985). Hiervan wordt een lineaire versie gebruikt waarvan de output door dubbele spatiale differentiatie wordt opgescherp. Deze contrastopscherping is via een onzekerheid relatie in tijd en frequentie gekoppeld aan een grotere groepsvertraging (*group delay*).

Vervolgens wordt de respons van het cochleamodel op sinustonen van verschillende frequenties besproken. Verschillende frequenties komen op verschillende posities maximaal tot expressie (hun karakteristieke positie), en omgekeerd heeft elke positie een karakteristieke frequentie. Dit wordt samengevat in een frequentie-plaatsrelatie. Met behulp van deze frequentie-plaatsrelatie en met behulp van de karakteristieke vorm van de sinusresponsen is het mogelijk om de energie- en frequentieontwikkeling van de individuele signaalcomponenten nauwkeurig te volgen. Omgekeerd is het mogelijk om het cochleogram van een quasiperiodieke bron te reconstrueren wanneer bekend is welke *ridges* aan de bron toegeschreven dienen te worden.

Hoofdstuk 4: The Time Normalized Correlogram

Hoofdstuk 4 behandelt de TNC als generalisatie van het cochleogram waaraan periodiciteit als extra dimensie is toegevoegd. Na een vergelijking van drie correlogramdefinities wordt het temporele gedrag van de TNC besproken. De TNC heeft twee stabiele basisvormen die steeds als superpositie voorkomen. Eén basisvorm is geassocieerd met periodieke signalen, de andere met aperiodieke signalen. De aperiodieke vorm wordt gedefinieerd door de impulsrespons van de TNC.

Vervolgens wordt ingegaan op de *Characteristic Period Correlation* (CPC), die ook een deelverzameling van de TNC is. De CPC meet de mate waarin basilaire membraan posities oscilleren met hun karakteristieke frequentie. Dit is een belangrijke indicatie voor het potentiële belang van een deel van het cochleogram oppervlak. Oppervlakken die niet met hun eigen karakteristieke

frequentie trillen worden gerekruteerd door signaalcomponenten die elders maximaal tot expressie komen.

Een bijzondere consequentie van de twee stabiele TNC basisvormen is dat signalen in termen van een periodiek en een aperiodiek deel kunnen worden beschreven. Dit biedt nieuwe mogelijkheden voor de analyse van complexe en variabele signalen die nader onderzocht dienen te worden.

Na een meer gedetailleerde behandeling van de vaststelling en de maximaal mogelijke nauwkeurigheid van de Locale Instantane Frequentie schatting, wordt het hoofdstuk afgesloten door een meer gedetailleerde analyse van de *Tuned Autocorrelation*. Na de behandeling van een correctie voor schattingsfouten in de periode en een correctie voor *group delay* effecten wordt het TAC schattingsproces in ruis besproken.

Hoofdstuk 5: Fundamental Period Estimation

Het schatten van een periodecontour is niet triviaal. Het is met name moeilijk om een betrouwbaar algoritme voor willekeurige ruizige omstandigheden te ontwikkelen. In het eerste deel van hoofdstuk 5 wordt het algoritme dat is geïntroduceerd in hoofdstuk 2 in meer detail besproken. In het tweede deel wordt een algoritme gepresenteerd dat geschikt is voor schone signalen en dat een variatie is van al bekende correllogramsommatiemethoden. Het verschil met deze methoden is dat er alleen gebruik gemaakt wordt van de informatie onder *ridges*. Dit algoritme was verantwoordelijk voor de periodecontouren voor het spraakherkenningsexperiment in hoofdstuk 2.

Hoofdstuk 6: Auditory Element Estimation

In hoofdstuk 6 wordt aangetoond dat de doelstelling van het onderzoek, zoals geformuleerd in hoofdstuk 1 inderdaad zijn gerealiseerd. Het hoofdstuk start met het schatten van zogenaamde *Auditory Elements*: cochleogramoppervlakken die op basis van overeenkomst met de daar tot expressie komende (fysische) eigenschappen aan één (doorgaans nog onbekende) bron kunnen worden toegeschreven. Hierbij worden een aantal criteria gebruikt die afzonderlijk vrij zwak zijn (d.w.z., dat er gemakkelijk aan voldaan kan worden), maar die in combinatie tot een betrouwbaar resultaat leiden.

Eén criterium is de vraag of de energie boven de waarde van een achtergrondmodel uitkomt. Een ander criterium stelt vast of er sprake is van voldoende toegevoerde energie. Een volgend criterium, op basis van de CPC, controleert of er sprake is van een locale frequentie die past bij de positie (locale dominan-

tie). Een laatste criterium is gebaseerd op de TAC en geeft aan welke delen een periodiciteit vertonen die overeenkomt met een vastgestelde periodecontour. Door deze criteria op verschillende manieren te combineren is het mogelijk om kwalitatief verschillende *auditory elements* te identificeren. Zo is het mogelijk om enerzijds individuele harmonischen en complexen van harmonischen (formanten) en anderzijds aperiodieke componenten (bijvoorbeeld een s-klink) te identificeren. Deze bijdragen zijn met behulp van de technieken in de vorige hoofdstukken verder te analyseren. De *auditory elements* blijken in veel gevallen, ook in zeer ruizige situaties (zoals 0 dB babbels-ruis), één op één te corresponderen met afzonderlijke signaalcomponenten.

In de rest van het hoofdstuk worden twee maten berekend die een indicatie geven van de overeenkomst tussen het originele ruizige signaal en de informatie die gerepresenteerd wordt in de *auditory elements*. In eerste instantie wordt een maat berekend op basis van de vergelijking van een cochleogram van het originele signaal en een cochleogram van een geresynthetiseerde versie van het signaal. Deze vergelijking laat een verbetering van de signaal-ruis-verhouding zien van maximaal 18 dB.

Bij de tweede en laatste vergelijking wordt de afstandsmaat gebaseerd op de *ridges* (die in veel gevallen één op één corresponderen met individuele signaalcomponenten). Doordat de *ridges* corresponderen met de meest ruisrobuuste delen van het cochleogram is hierbij als functie van een afnemende signaal-ruis-verhouding in eerste instantie veel minder sprake van een degradatie van het bewerkte ruizige signaal. Bij een signaal-ruis-verhouding van 0 dB (evenveel signaal als ruis energie) is de correlatie tussen het schone en het bewerkte signaal nog steeds rond de 0.95 en is er sprake van een verbetering van de signaal-ruis-verhouding van rond de 25 dB.

Hiermee wordt het waarde van de in hoofdstuk 1 geformuleerde aanpak bevestigd.

Hoofdstuk 7: Overview and Discussion

In het laatste hoofdstuk wordt CPSP samengevat, worden de formules algemener geformuleerd en wordt ingegaan op een aantal mogelijk uitbreidingen. Het hoofdstuk vervolgt met een schetsmatig voorstel van een spraakherkenningsaanpak die vrij nauw aansluit bij moderne ASR technieken, maar waarin de spraak-in-ruis-paradox vermeden kan worden. Het proefschrift wordt af-

gesloten met een korte analyse van de relevantie van CPSP voor het onderzoek naar menselijke spraakherkenning en een overzicht van de bijzondere eigenschappen van CPSP.