Automatische analyse van pathologische spraak

Automatic Analysis of Pathological Speech

Catherine Middag

Promotor: prof. dr. ir. J.-P. Martens
Proefschrift ingediend tot het behalen van de graad van
Doctor in de Ingenieurswetenschappen

Vakgroep Elektronica en Informatiesystemen
Voorzitter: prof. dr. ir. J. Van Campenhout
Faculteit Ingenieurswetenschappen en Architectuur
Academiejaar 2012 - 2013

UNIVERSITEIT
GENT

ISBN 978-90-8578-536-1
NUR 981
Wettelijk depot: D/2012/10.500/62

Table of Contents

List of Figures

List of Tables

List of Acronyms

A

ASR	Automatic Speech Recognizer
ARPA	Advanced Research Projects Agency
ARPABET	ARPA phonetic alphabet
ALF-PLF	alignment-free phonological feature
ALF-PMF	alignment-free phonetic feature

C

CGN	Corpus Gesproken Nederlands
CoGeN	Corpus Gesproken Nederlands
CVC	Consonant-Vowel-Consonant
CFDA	Computerized Frenchay Dysarthria Assessment
CV	cross-validation
CD	context-dependent
CI	context-independent
COPAS	Dutch Corpus of Pathological Speech
CMS	Cepstral Mean Subtraction
CD-PLF	context-dependent phonological feature
CCRT	Concurrent Chemo-Radiotherapy

D

DIA	Dutch Intelligibility Assessment
DTW	Dynamic Time Warping
DU	Dutch

E

E-speech	Esophageal speech
EL-speech	Electrolaryngeal speech
ELR	Ensemble Linear Regression
EBP	error back-propagation
EM	Expectation-Maximization

F

FDA	Frenchay Dysarthria Assessment
FL	Flemish

G

GOF	Goodness of Fit
GMM	Gaussian Mixture Model

H

HMM	Hidden Markov Model

I

IPA	International Phonetic Alphabet
IPM	Intelligibility Prediction Model

L

LDA	Linear Discriminant Analysis
LLR	Log Likelihood Ratio
LME	Lehrstuhl für Mustererkennung

M

ML	Machine Learning
MLP	Multi Layer Perceptron
MFCC	Mel-frequency cepstral coefficient
MAP	Maximum A Posteriori

N

NSVO	Nederlandstalig Spraakverstaanbaarheidsonderzoek
NKI-UVAFON	Nederlands Kankerinstituut - instituut voor fonetiek aan de Universiteit Amsterdam

P

PE	pharyngo-esophageal
PE-segment	pharyngo-esophageal segment
PI	Phoneme Intelligibility
PEAKS	Program for Evaluating All Kinds of Speech disorders
PLAKKS	Psycho-Linguistische Analyse Kindlicher Sprech-Störungen
PCC	Pearson Correlation Coefficient
PCA	Principal Component Analysis
PFD	phonological feature detector
PLF	phonological feature
PMF	phonemic feature

R

RSI	Running Speech Intelligibility
RF	Random Forests
RBF	radial basis function
RMSE	root mean squared error

S

SAMPA	Speech Assessment Methods Phonetic Alphabet
SPACE	Speech Algorithms for Clinical and Educational applications
SVM	Support Vector Machine
SVR	Support Vector Regression

T

TE-speech	tracheo-esophageal speech
TM	Text Marloes

U

UZA	Universitair Ziekenhuis Antwerpen
UBM	Universal Background Model

V

VOT	voice onset time

W

WA	Word Accuracy
WAR	Word Accuracy Rate

Y

YAPA	Yet Another Phonetic Alphabet

List of Symbols

F_0	fundamental frequency
X	observation vector
$\mathbf{X}$	matrix representing a dataset of examples or sequence of observations
Y	vector containing the targeted outputs of a model
$\boldsymbol{\beta}$	vector containing the regression coefficients of a model
$\boldsymbol{\Sigma}$	covariance matrix
$\boldsymbol{\mu}$	vector of means
ξ	vector of slack variables in a Support Vector Machine
$s(t)$	speech signal
$\mathbf{W}$	a sequence of words
$\mathbf{F}$	a sequence of phone(me)s
$P(\mathbf{F}\|\mathbf{W})$	posterior probability of $\mathbf{F}$ given $\mathbf{W}$
s_t	model state visited at time t
$\mathcal{S}$	the set of all possible states
$\mathbf{S}$	sequence of states
A	vector of phonological class properties
A_{ci}	canonical value of phonological class property A_i

Samenvatting

Verbale communicatie is niet meer weg te denken uit ons dagelijks leven en wordt maar al te vaak als vanzelfsprekend beschouwd. Nochtans ondervinden mensen met spraakstoornissen hier zowel praktisch als sociaal grote problemen door. Daarom is het belangrijk dat deze mensen kunnen behandeld en opgevolgd worden door logopedisten.

De ernst van een spraakstoornis wordt vaak uitgedrukt in termen van de spraakverstaanbaarheid. Dit begrip heeft een ruime betekenis, maar kan hier worden gedefinieerd als de mate waarin een luisteraar in staat is de boodschap van de betreffende spreker te "ontcijferen". In de dagelijkse praktijk wordt spraakverstaanbaarheid gemeten d.m.v. een perceptuele test, waarbij de spreker een bepaalde tekst of woordenlijst voorleest, en waarbij de verstaanbaarheid van de spraak beoordeeld wordt door een logopedist(e). De resultaten van deze werkwijze zijn echter niet altijd even betrouwbaar. Ze zijn immers van nature subjectief, vermits de logopedist(e) de patiënt kan kennen (spreekstijl, stem) en vermits het herhaaldelijk afnemen van dezelfde test ertoe leidt dat de logopedist(e) de spreker beter zal begrijpen dan een luisteraar die niet met de test bekend is.

Om deze redenen zou het interessant zijn om te kunnen beschikken over een altijd objectieve luisteraar die men zou kunnen creëren door gebruik te maken van automatische methoden ter bepaling van spraakverstaanbaarheid. De afgelopen jaren zijn er reeds enkele programma's ontwikkeld die dit doel voor ogen hebben, maar die programma's zijn nog onvoldoende getest om met succes hun intrede te kunnen doen in de klinische praktijk.

Van 2005 tot 2009 voerden de universiteiten van Brussel, Gent en Leuven (België) samen met het universitair ziekenhuis van Antwerpen (UZA) een IWT-SBO-project uit met als titel: 'Speech Algorithms for Clinical and Educational applications' (SPACE). In het kader van dit project ontwikkelde ik een programma voor het automatisch bepalen van de spraakverstaanbaarheid van Vlaamse pathologische sprekers op basis van de opnames die normaal in het kader van een standaard perceptuele test worden gemaakt.

Spraakverstaanbaarheid op foneemniveau

In eerste instantie onderzocht ik of het mogelijk is de bestaande perceptuele verstaanbaarheidstest, het "Nederlandstalig Spraakverstaanbaarheidsonderzoek" of afgekort NSVO genaamd, te automatiseren. Deze test, die in het Engels als "Dutch

Intelligibility Assessment" (DIA) bekend staat, onderzoekt de verstaanbaarheid van alle Vlaamse fonemen op basis van uitspraken van 50 geïsoleerde woorden waarvan er een groot deel geen betekenis hebben. Per woord dient de logopedist(e) één foneem te identificeren (b.v. de finale consonant). De verstaanbaarheid wordt dan bepaald als het percentage correct geïdentificeerde fonemen. De ontwikkelaars van de DIA (logopedisten van het UZA) bezorgden mij de opnames van een 200-tal pathologische en een honderdtal niet-pathologische sprekers die deze test voorlazen, alsook nog wat ander tekstmateriaal. Samen met de bijhorende verstaanbaarheidsscores werden deze opnames opgenomen in het Corpus voor Pathologische Spraak (COPAS) dat als basis voor mijn onderzoek heeft gefungeerd.

Ik onderzocht verschillende mogelijke pistes om op basis van COPAS een systeem te bouwen voor het automatisch bepalen van spraakverstaanbaarheid. Ruwweg werkt een dergelijk systeem in drie stappen: een **voorverwerking** die een spectro-temporele voorstelling van het akoestisch signaal oplevert, een **sprekerkenmerkenextractor** die deze voorstelling analyseert en er globale kenmerken uit afleidt die de spraak van de spreker op een compacte manier beschrijven, en een **spraakverstaanbaarheidsvoorspelling** die op basis van deze kenmerken de spraakverstaanbaarheid berekent. Een van de belangrijkste uitdagingen van mijn onderzoek was het afleiden van interessante sprekerkenmerken die genoeg informatie bevatten om er verstaanbaarheid te kunnen uit afleiden.

Oorspronkelijk bevatten al mijn systemen een automatische spraakherkenner die dezelfde taak uitvoerde als de logopedist(e) in een perceptuele test: naar de 50 woorden luisteren en het geteste doelfoneem trachten in te vullen. Door de resultaten van de herkenner te vergelijken met wat de spreker gevraagd werd te lezen, kan men dan zogenaamde **foneemnauwkeurigheid** bepalen.

Omdat deze foneemnauwkeurigheid onvoldoende met de perceptuele waarde correleerde werden alternatieve systemen ontwikkeld waarin de spraakherkenner de spraak alleen maar dient op te lijnen met de gekend veronderstelde uitgesproken tekst. Met behulp van een competitieve spraakherkenner, gebaseerd op de traditionele fonetische modellen voor Vlaamse (normale) spraak, werd het pathologische spraaksignaal opgelijnd met de fonetische transcriptie van de gelezen tekst (een woord). Uit die oplijning werden dan **fonemische kenmerken** afgeleid, die beschrijven hoe goed de waargenomen uitspraken van de Vlaamse fonemen in de woorduitspraken van een spreker gemiddeld genomen worden 'herkend' door de (normale) foneemmodellen van de spraakherkenner.

Vermits spraakverstaanbaarheid veel te maken heeft met articulatie, onderzocht ik vervolgens of fonologische (articulatorische) modellen voor spraak een meerwaarde zouden kunnen bieden t.o.v. de traditionele fonemische modellen. Daartoe werd het pathologische spraaksignaal nogmaals opgelijnd met de fonetische transcriptie van de doelwoorden, maar deze keer werd daarbij een spraakherkenner gebruikt die met fonologische modellen voor (normale) Vlaamse spraak werkt. Hieruit werden dan de **fonologische kenmerken** afgeleid, die beschrijven hoe goed de fonologische eigenschappen van de waargenomen uitspraken van

Vlaamse fonemen in de woorduitspraken van een spreker gemiddeld genomen worden 'herkend' door de (normale) fonologische modellen van de spraakherkenner.

Ik onderzocht het potentieel van al deze sprekerkenmerken als voorspellers van spraakverstaanbaarheid door ze als inputs van diverse verstaanbaarheidsmodellen te gebruiken. Dit bracht aan het licht dat een combinatie van foneemnauwkeurigheid en fonologische kenmerken m.b.v. een Support Vector Regressie kon omgezet worden tot betrouwbare spraakverstaanbaarheidsscores die zeer goed correleren met de perceptuele beoordelingen (Pearson Correlatiecoefficiënt van meer dan 0.80). Een andere interessante combinatie was deze van fonologische en fonemische kenmerken. Zij leidt tot scores die niet significant minder dan de uitgangen van het beste model met de perceptuele scores correleren. Bovendien geven ze meer inzicht in de onderliggende articulatorische problemen van de geteste spreker.

Behalve de zonet genoemde algemene modellen, werden ook pathologie-specifieke modellen ontwikkeld. Deze leiden meestal tot hogere correlaties (tot 0.96) tussen de perceptuele en de automatische verstaanbaarheidsscores.

Door te onderzoeken welke kenmerken belangrijk zijn voor de ontwikkeling van een verstaanbaarheidsmodel, vond ik dat de kenmerken die frequent door de machinale leermethodes geselecteerd werden kunnen gelinkt worden aan gekende articulatieproblemen bij specifieke types pathologische sprekers.

Verstaanbaarheid op zinsniveau

Vermits de NSVO test voor de helft uit nonsenswoorden bestaat, kunnen mensen beginnen twijfelen aan het woord dat ze dienen te lezen, en lezen ze een woord dat ze kennen. Ze maken dus leesfouten die verkeerdelijk tot een reductie van de verstaanbaarheidsscore leiden. Om dat probleem te omzeilen zocht ik naar een alternatieve manier om de verstaanbaarheid op foneemniveau af te leiden uit opnames van normale lopende spraak. Gebruik makende van dezelfde methodologie als hierboven beschreven, ontwikkelde ik een verstaanbaarheidsmodel dat werkt met fonologische en fonemische sprekerkenmerken die zijn afgeleid door oplijning van de pathologische spraakuitingen met de zinnen die ze voorstellen. Dit leidde opnieuw tot betrouwbare verstaanbaarheidsscores.

Hoewel deze methodologie op basis van de oplijning van lopende spraak goed schijnt te werken, worden de resultaten toch nog steeds negatief beïnvloed door aarzelingen en leesfouten in de opnames. Daarom ontwikkelde ik een **oplijningsvrije fonologische en fonetische karakterisering** van de spreker die kan worden afgeleid zonder gebruik te moeten maken van kennis over de gelezen tekst. Ik kon aantonen dat een verstaanbaarheidsmodel gebaseerd op deze nieuwe kenmerken eveneens tot zeer goede resultaten leidt.

Vermits de afgeleide sprekerkenmerken tot goede resultaten leidden op Vlaamse data, werd vervolgens onderzocht of ze ook bruikbaar zijn voor andere talen of dialecten. Met dit doel voor ogen werd samengewerkt met de universiteit van Er-

langen (Duitsland) en het Nederlands Kankerinstituut en de Universiteit van Amsterdam. Via deze samenwerkingen kon ik aantonen dat de voorgestelde methodologie voor de berekening verstaanbaarheid op zinsniveau ook in die talen (Duits) en regionale varianten (Noord-Nederlands) tot scores leidt die nagenoeg even betrouwbaar zijn als perceptuele scores.

Verdere analyse van pathologische spraak

Nu dat aangetoond is dat verstaanbaarheid betrouwbaar kan bepaald worden met behulp van de door mij ontwikkelde sprekerkenmerksets, is het tijd om te onderzoeken of deze sprekerkenmerken meer in zich hebben. Vermits ze naar specifieke dimensies verwijzen (fonologisch, fonemisch of fonetisch), kunnen ze misschien ook specifieke problemen voorspellen die betrekking hebben op deze dimensies, en dus m.a.w. de onderliggende oorzaken van een lage verstaanbaarheid naar boven brengen. Een eerste onderzoek op basis van COPAS leidde niet tot bevredigende resultaten, in de eerste plaats omdat dit corpus weinig of geen betrouwbare informatie over de manier van spreken van de patiënten bevat.

Anderzijds bleek het wel mogelijk om een groep sprekers met een welbepaalde pathologie op basis van een beperkt aantal goed gekozen sprekerkenmerken te onderscheiden van de niet-pathologische sprekers. Doordat dit reeds goed bleek te werken met slechts twee kenmerken waren de resultaten van een dergelijke analyse gemakkelijk te visualiseren. Uit de experimenten bleek dat vooral de oplijningsvrije kenmerken hier goed scoorden, en dat de de meest onderscheidende kenmerken konden gekoppeld worden aan eigenschappen van de onderzochte pathologieën.

Door samenwerking met het Nederlands Kankerinstituut en de Universiteit van Amsterdam kon ik experimenteren op een corpus waarin alle pathologische sprekers beoordeeld waren door 13 luisteraars. Verder werden ook verschillende aspecten van hun spraak beoordeeld: verstaanbaarheid, fonatie, articulatie, dialect, enzovoorts. Gebruik makende van eenzelfde methodologie als voor de ontwikkeling van verstaanbaarheidsmodellen, kon ik modellen ontwikkelen voor het voorspellen van fonatiekwaliteit, de articulatiekwaliteit en het accentgehalte van de spraak. Deze modellen bleken minstens even betrouwbaar als een gemiddelde luisteraar.

De DIA-tool

Mijn onderzoek leidde tot de ontwikkeling van de online DIA tool. Deze tool is gebruiksvriendelijk en gemakkelijk toegankelijk voor logopedisten. Men heeft enkel een PC of laptop nodig met een web browser, een head set, een geluidskaart en een Java runtime environment. Het programma werkt in een client/server omgeving en kan zowel online als offline gebruikt worden. De tool laat toe de oorspronkelijke NSVO woordtest op te nemen en er zowel een perceptuele test als een automatische analyse op uit te voeren. De automatische analyse leidt dan tot een

rapport dat de huidige verstaanbaarheid van de patiënt bevat alsook enkele figuren waarin de positie van de spreker t.o.v. andere pathologische en normale sprekers wordt aangegeven. Binnenkort zal ook de automatische test op zinsniveau in de tool worden opgenomen.

Summary

Effective verbal communication is an essential aspect of daily life and is often taken for granted. It presents a major bottleneck though for people experiencing speech disorders. Disordered (or pathological) speech can be the consequence of a plurality of causes, and the assessment, treatment and monitoring of causes have been receiving growing attention in the biomedical field.

A widely used measure of the severity of a speech disorder is speech intelligibility, loosely defined as the ease with which a listener is able to lexically decode the utterances of a speaker. In the clinical setting, measures of speech intelligibility for text level stimuli are often acquired by means of a perceptual test, but the results of such a test are anticipated to be subjective and influenced by the listener's familiarity with both the patient's voice and the prompted text.

For these reasons it would be interesting if one could have access to an ever objective assistant to score the intelligibility of the speech. This calls for the development of objective automatic methods for intelligibility assessment. During the past decade, a few attempts were made to develop such methods, but thus far these methods were insufficiently tested for getting widely accepted in clinical practice.

From 2005 to 2009, the universities of Brussels, Ghent and Leuven (Belgium) and the University Hospital of Antwerp participated in an IWT-SBO-project with the title 'Speech Algorithms for Clinical and Educational applications' (SPACE). In the course of this project I developed a program for the automatic computation of speech intelligibility of Flemish pathological speakers on the basis of recordings that are normally made as part of a standard perceptual test.

Phoneme Intelligibility

As a first step, I investigated whether it was possible to automate the Flemish phoneme intelligibility test, called the "Nederlandstalig Spraakverstaanbaarheidsonderzoek" (NSVO). This test, referred to as "Dutch Intelligibility Assessment" (DIA) in English literature, examines the intelligibility of all Flemish phonemes on the basis of 50 isolated monosyllabic word utterances, a large part of which are actually meaningless. Per word the speech therapist has to identify one phoneme (e.g. the final consonant). Intelligibility is then defined as the percentage of correctly identified phonemes. The developers of the DIA test (UZA) kindly provided recordings of about 320 speakers, amongst which about 200 pathological speakers. For each speaker, recordings of the DIA test and of some other tests (e.g. a para-

graph) were available. The recordings, together with their perceptual intelligibility scores, were assembled in COPAS, a corpus of pathological and normal speech.

I used COPAS to develop and investigate several methods for the automatic prediction of speech intelligibility.

Roughly speaking, such a system operates in three steps: a **preprocessing** stage which produces a spectro-temporal representation of the acoustic signal, a **speaker feature extraction** which analyses this representation and retrieves global features from it to construct a compact characterization of the speech of the tested speaker, and an **intelligibility prediction** which computes an intelligibility score on the basis of these speaker features. Deriving interesting speaker features which carry enough information about individual intelligibility problems was one of the main challenges of my research.

Originally, all my systems comprised an automatic speech recognizer (ASR) that performed exactly the same task the speech therapist has to perform in a perceptual task, namely, listen to the 50 words and identify the tested phoneme. By comparing the outputs of the speech recognizer with the prompts that were given to the speaker, one obtains a **phoneme accuracy**.

Because the computed phoneme intelligibility did not sufficiently correlate with its perceptual value, alternative systems were developed employing the speech recognizer just to align the speech with the prompted text (a word in this case). By means of a state-of-the-art ASR, working with traditional context-dependent phonemic models of Flemish (normal) speech, the pathological speaker's utterance was aligned with the phonemic transcription of the prompted text. From this alignment a number of global features were derived which describe how well the observed pronunciations of the Flemish phonemes were 'recognized' by the phoneme models inside the recognizer.

As speech intelligibility is closely related to articulation, I investigated whether phonological models offer an added value over the the traditional phonetic models. To establish this, the pathological speech is aligned again with the prompted text, this time with an ASR employing phonological models for (normal) Flemish speech. From this alignment, **phonological speaker features** were derived to describe how well on average the observed pronunciations of Flemish phonemes in word utterances of a speaker are 'recognized' by the normal phonological models of the ASR.

I investigated the potential of all these speaker feature sets as predictors of speech intelligibility by employing them as inputs to diverse intelligibility prediction models. This investigation revealed that by means of Support Vector Regression it is possible to convert the combination of phoneme accuracy and phonological features to an intelligibility score that correlates very well with the perceptual ratings (Pearson Correlation Coefficient higher than 0.80). Another interesting combination was that of the phonological and the phonemic features. The scores emerging from that combination correlate nearly as well to the perceptual ratings than those emerging from the best model. Moreover, they offer more insight in the underlying articulatory problems of the tested speaker.

Next to the above general models, I also created pathology-specific models.

These models generally yield higher correlations (up to 0.96) between perceptual and automatic intelligibility scores.

By analyzing which features are important for the development of an intelligibility prediction model, I found that the features that are frequently selected by the machine learning methods can be linked to known articulatory problems in specific pathological speaker populations.

Running Speech Intelligibility

Since the DIA test material consists for more than 50% of nonsense words, people may start to doubt about what they have to read, and substitute the prompted word by a common word they know. This leads to reading errors which on their turn lead to a reduction of the intelligibility score. To circumvent this problem I have searched for an alternative way of deriving phoneme intelligibility from recordings of running speech. Using the same methodology as before, I developed an intelligibility prediction model that employs phonological and phonemic speaker features that originate from alignments of speech utterances with the sentences they represent. Again, this approach yielded reliable intelligibility scores.

Although the alignment-based methodology seems to perform well, the results continue to be affected by hesitations and reading errors in the recordings. That is why I also developed an **alignment-free phonological and phonetic speaker characterization** that can be calculated without having to employ any knowledge of the prompted text. I was able to demonstrate that an intelligibility model based on the new speaker features also leads to reliable results.

As the derived speaker features worked well for Flemish data, I also investigated their usability for other languages and regional variants/dialects. With this goal in mind a collaboration was set up with the Chair of Pattern Recognition of the University of Erlangen (Germany) and the Dutch Cancer Institute and the Institute of Phonetics of Amsterdam University (Netherlands). Through these collaborations I succeeded in demonstrating that the proposed methodology for the prediction of running speech intelligibility (RSI) at the sentence level also leads to reliable scores in the tested languages (German) and regional variants (Northern Dutch). The computed scores turn out to be about as reliable as the perceptual scores of a single human rater.

Further analysis of pathological speech

Having shown that intelligibility can be predicted in a reliable way on the basis of the speaker features I developed, I investigated whether these features have more potential than that. Since the features point to specific dimensions (phonological, phonemic or phonetic), they might be able to predict specific problems that relate to these dimensions. Consequently, they may be able to reveal the underlying causes of a low intelligibility. An initial investigation using COPAS did not lead

to the desired results, mainly because this corpus comprises little or no reliable detailed information about the way patients speak.

On the other hand, it appeared viable to separate a group of speakers with a specific pathology from the normal speakers on the basis of a small number of well chosen speaker features. As two features seemed to be enough to get good results, the results were also easy to visualize. Experiments showed that alignment-free features scored very well here, and that the most distinctive features could be linked to properties of the considered pathology.

By collaborating with the Dutch Cancer Institute and the Institute of Phonetics of Amsterdam University I had the opportunity to work on the NKI-CCRT database. In this database, all pathological speakers were judged by 13 human raters on different aspects such as intelligibility, articulation, phonation, accent, etc. Using the same methodology as before, I was able to develop phonation quality, articulation quality and accent gravity prediction models that seem to be as reliable as the average human rater.

The DIA tool

My research has lead to the development of the DIA tool. The tool is easy to use and easy to access by speech therapists. All that is needed is a PC or laptop with a web browser, a head set, a sound card, and an up-to-date Java runtime environment. The tool works in a client/server environment and can be used both in an on-line and an off-line mode. The tool permits to record the original DIA word test and to conduct a perceptual test as well as an automated analysis. The automatic analysis produces a report describing the patient's current intelligibility and some figures showing his position against other pathological speakers and against the normal speakers. Soon, the automatic RSI test will also be incorporated in the DIA tool.

1 General Introduction

1.1 Intelligibility

Verbal communication is getting increasingly important in our society. As we need the possibility to interact with other people for almost all of our basic needs, it has become an essential part of our lives. People with speech disorders therefore suffer from great functional as well as social discomfort as they are deprived from one of the primary communication forms. This explains the growing need for speech therapy as a follow-up of these patients in order to assess and improve their vocal and pronunciation skills.

During speech rehabilitation, it is important that the speech therapist regularly measures the current state of the patient in terms of type, progression and severity of the pathology. This monitoring can be useful to determine the right and personalized therapy for every individual patient. One of the most popular and important measures for severity of a speech disorder is intelligibility.

Intelligibility is an overall measure which can be defined as the degree to which the acoustic realization of one's speech can be understood [1]. This does not involve semantic or syntactic context, neither visual aspects of communication like e.g. gestures [2]. A person's intelligibility is purely determined by the performance of his/her speech production system.

Traditionally, the patient's intelligibility is measured using a perceptual test: the speech therapist listens to and rates the patient's utterance. This method is by definition subjective in nature, affecting the reliability of the outcome. Several

factors can bias the listener, such as familiarity with the speaker or the speaker's accent, background and type of disorder. Similarly, familiarity with the used test material also induces a positive bias. In tests using real words, linguistic information can also play a role, as the speech therapist can guess the right word by using the context or language knowledge. Some intelligibility tests try to circumvent this problem by using large sets of test items combined with random selection or by working with non-existing words or meaningless sentences to rule out the usage of linguistic knowledge. However, this influences the naturalness of the test items in a negative way, leading to reading mistakes and hesitations at the speaker's side [2]. This trade-off between naturalness of the test material and predictability for the speech therapist is difficult to keep balanced. And even when this would not be a problem, the familiarity with the (type of pathological) speaker still remains an issue. Speech therapists who work with some type of pathological speakers automatically know their typical errors and, when therapy evolves, they get familiar with their patients' speech. As a logical consequence, it is impossible to assess ones intelligibility without previous knowledge/bias. Therefore, it could be interesting to create an assistant which always stays objective. This calls for the development of objective automatic methods for intelligibility assessment since the use of computer-assisted measurements could solve the subjectivity-issues.
During the last decade, only a few attempts were made to develop an automated intelligibility assessment tool [3,4]. From 2005 to 2009, the universities of Brussels, Antwerp, Ghent and Leuven (Belgium) and the University Hospital of Antwerp participated to the the IWT-project 'Speech Algorithms for Clinical and Educational applications' (SPACE) [5]. As one of the members of this research project, I developed an automatic intelligibility assessment tool for Flemish pathological speakers.

Of course, introduction of technology always needs proof of reliability. Therefore, an automatic method should prove to provide scores which are as reliable as those of a panel of speech therapists which are not particularly familiar with the patient. Therefore, I started from the Flemish phoneme intelligibility test, called the "Nederlandstalig Spraakverstaanbaarheidsonderzoek" (NSVO) [6], referred to as "Dutch Intelligibility Assessment" (DIA) in English literature. This test was developed by the University Hospital of Antwerp (UZA) and consists of 50 isolated short words, mostly nonsense words. Recordings of 319 speakers, amongst which 197 pathological, performing this test, were kindly provided by the UZA, together with their perceptual intelligibility score according to the DIA. This formed the base for my research, which eventually lead to the online and freely available DIA-tool [1].

As mentioned above, the need for objectivity in perceptual tests often leads to the use of unnatural speech material like nonsense words or sentences, leading

[1] http://diaweb.elis.ugent.be/

to reading mistakes and confusions. As the perceptual DIA-test uses nonsense words, a possible improvement for the DIA-tool could be to only use existing words or, even closer to daily life, use sentences or paragraphs. After automating the perceptual DIA test, I investigated whether I could predict intelligibility from running speech.

Finally, intelligibility is, as much as it is used, just one number. It does not inform the speech therapist about the underlying reasons for a high or low score, nor does it point to specific articulatory problems. In my opinion, an intelligibility test should be able to present more than just one number. A more profound articulatory analysis can help the speech therapist to determine the right personal pathology for every patient, and is therefore very valuable. My latest research is devoted to this topic.

1.2 Terminology

At the start of this dissertation, it is useful to elucidate some frequently used terms in the field of speech science.

Phoneme : the smallest contrastive unit in the sound system of a particular language. It serves to distinguish between meanings of words. The phonemes of a language constitute the minimal set of symbols needed to describe the pronunciations of all words in that particular language. All Flemish phonemes can be found in Appendix A.

Allophone : phonemes can be pronounced in different ways while still carrying the same meaning. Those different pronunciations are called allophones. In Flemish, tongue-tip /r/ and uvular /R/ are different allophones of the same phoneme /r/.

Phone : the acoustic realizations of some of the allophones typically consist of two successive phases, like e.g. the pressure build-up and the pressure release phase of a /p/. Therefore, one has also introduced sub-phonemic units representing those phases. Together with the allophones not needing such a split, they constitute a set of phones that can be used for an acoustic-phonetic description of the speech.

Orthographic transcription : the written record of speech. Orthographic refers to the use of the standard alphabet.

Phonemic transcription : this form of transcription uses a sequence of phonemes to describe the uttered speech.

Phonetic transcription : this form of transcription uses a sequence of phones to describe the uttered speech.

Phonetic alphabet : The International Phonetic Association (IPA) [7] has developed an alphabet, called the International Phonetic Alphabet, for describing as accurately as possible all phones in all possible languages. This phonetic alphabet is internationally used. As the symbols used in this alphabet are not the standard symbols found on computer keyboards, there exist a number of encodings representing a subset of the IPA that can be used to describe the speech in one language. Examples of such encodings are: SAMPA (Speech Assessment Methods Phonetic Alphabet, [8]), YAPA (Yet Another Phonetic Alphabet) and CGN (Corpus Gesproken Nederlands - see [9]). All Flemish phonemes can be found in Appendix A with their notation in SAMPA, YAPA, CGN and IPA and a Flemish word in which they appear. In this dissertation, I will use the CGN phonetic alphabet for Flemish (51 phones for 45 phonemes, extended with four extra symbols for annotating short and long silences (/#/ and /##/), glottis closure (/!/) and unknown (/?/)) and the ARPABET alphabet for American English (48 phonemes, 58 phones and some pause symbols) [10], which can also be found in Appendix A. Obviously, the ARPABET comprises much more allophones than the CGN alphabet.

Articulatory characteristics of speech are characteristics stemming from direct measurements of articulatory movements (e.g. by means of an articulograph).

Phonological characteristics of speech are characteristics which intend to describe articulatory phenomena, although they are derived from the waveform [11].

Phonetic characteristics of speech describe the characteristics of speech in terms of its phones.

Phonemic characteristics of speech describe the characteristics of speech in terms of its phonemes.

1.3 Outline

As described in Section 1.1, this dissertation assembles my research towards an automatic intelligibility assessment tool. Before describing my methodology, I will briefly describe the human speech production system in Chapter 2 and some problems that can arise in this system, leading to speech disorders, in Chapter 3. Some common evaluation strategies for these speech disorders will be discussed in Chapter 4.

To construct automatic models for intelligibility prediction, two important issues need to be discussed: the databases and the modeling techniques. Modeling techniques, and, more general, the machine learning basics I used, will briefly

be described in Chapter 5. The databases I used for model development will be described in Chapter 6.

Having described all necessary medical and technical basics, Chapter 7 covers my research in the automation of the perceptual DIA test. Chapter 8 broadens the use of automatic intelligibility assessment tools towards running speech and Chapter 9 explores the prediction of more aspects of pathological speech.

Since my research lead to the online DIA-tool, this tool will be described in Chapter 10.

Finally, Chapter 11 ends this dissertation with the main conclusions of my work and some possible directions for future work.

1.4 Contributions

My research contributed to the field of speech analysis for applications which have to cope with disordered speech. First of all, I investigated whether phonological models for Flemish speech offer more potential for assessing articulation problems than the traditional phonetic models typically used for speech recognition.

I have proposed two novel approaches that can predict the intelligibility of disordered speech. The first approach is an alignment-based approach which uses phonological or phonetic models to align the speech with its phonetic transcription, and which derives from that alignment a number of global features that constitute the phonological or phonetic characterization of the speaker. This characterization is then supplied to an intelligibility prediction model that predicts the intelligibility of that speaker.

Although the alignment-based approach yielded very good results, these results are negatively affected by the presence of hesitations and reading errors in the speech recordings. To circumvent this problem, I have created an alignment-free phonological and phonetic characterization of the speaker which can be computed without taking the read text into account.

The alignment-based and alignment-free speaker characterizations were then investigated in a number of collaborations with foreign institutes (Chair of Pattern Recognition of the University of Erlangen and the Netherlands Cancer Institute and the Institute of Phonetics of Amsterdam University). From these studies I could conclude that the proposed methodology for the prediction of intelligibility also yields close-to-human performance in scenarios involving multiple languages or accents and multiple speech modes (e.g. isolated words and running speech). In a similar vein I established that the proposed methodology also yielded good performance for the assessment of some articulation and phonation problems.

The main results of my research are published in 3 international journal papers [12–14] and in the proceedings of 5 international conferences [15–19].

2
Speech Production

The ability to express oneself verbally is used so much in everyday life that we often take it for granted. We do not realize how complex the process of speech production is and how many body parts have to function and cooperate correctly to even transfer a simple message. Thoughts must be translated into linguistic representations (sequence of phonemes) in the brain. This sequence of phonemes leads to a sequence of motor commands to the vocal organs. From the brain's motor center, every single of these commands is sent to the lungs for creating the right amount of air pressure, to the vocal tract to create the right air vibration (called phonation) and then to the oral cavities to create the right resonance and articulation [20]. In this section, I will briefly describe every of these sub-processes.

2.1 Speech motor control

Every single action in the speech process is programmed, coordinated and directed by the motor part of the nervous system. This consists of the central nervous system, consisting of the brain and the spinal cord, and the peripheral nervous system being the collection of nerves connecting all muscles to the central nervous system. While the peripheral nervous system is more used for reflexes, the central system is the one planning and consciously executing tasks [20]. In the next sections, I will elucidate the speech motor part of both the peripheral and the central nervous system.

2.1.1 Central motor system

Several parts of the central motor system are used in the speech production process. **The premotor cortex** (see figure 2.1), which is part of the frontal cortex, takes care of the planning of volitional movements. **The motor cortex** (see figure 2.1), situated next to the premotor cortex [21], activates and initializes the volitional movements. It is organized in an inverted body scheme, which means that every part of this area is responsible for transmitting motor impulses to a specific muscle in the contralateral part of the body [22]. **The pyramidal system** transports motor pulses from the cortex to the brainstem and spinal cord. It carries the voluntary, fine movements and is also organized in an inverted body scheme. **The extrapyramidal system** also makes part of the neural pathways from the cortex to the brainstem and spinal cord. Unlike the pyramidal system, it carries the automatic, non volitional aspects of the movements, like e.g. the initiation of walking or speaking, transmitted by the pyramidal system. Moreover, two important neurotransmitters are produced in this area: dopamine and acetylcholine, regulating the speed of the initiation of movements and the muscle tonus [20, 22]. **The cerebellum**, located at the base of the brain, monitors the precision of the movements by comparing the initial motor signal to the execution of the movement and - if necessary - correcting it. The cerebellum is important for coordination, smoothing and precision of movements.

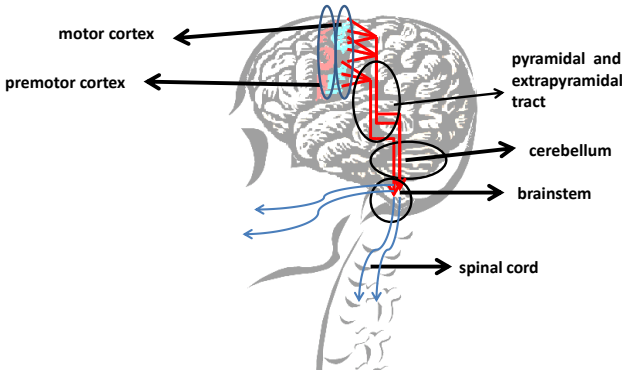

Figure 2.1: Motor speech pathway. Starting in the motor cortex, commands are transported through the pyramidal and extrapyramidal tract and passed to the brainstem where the necessary nerves are activated. After [20, 21].

2.1.2 Peripheral motor system

The peripheral motor system consists of a collection of cranial and spinal nerves, of which five are important for speech. They each control the movements of a specific part of head and neck. One nerve, the *nervus vagus*, is very important because it controls the velar, pharyngeal and laryngeal movements as well as the organs in the chest, such as breathing [20].

2.2 Vibration

After the commands of the nervous system have been distributed over the speech organs, the first step in the speech production process is the creation of a pressure wave. After inhalation, the air is pushed out of the lungs through the trachea to the larynx. This last part, illustrated in figure 2.2, is the most heavily innervated sensory structure in the body [23]. It is used for swallowing, protection of the airways and - very important - phonation. The cartilaginous framework of the larynx has two horizontal folds of soft tissue in the passage of air, called the vocal folds [24]. Right above the vocal folds, another pair of folds, called the vestibular folds or "false vocal folds" is located. The vocal folds divide the airway in a

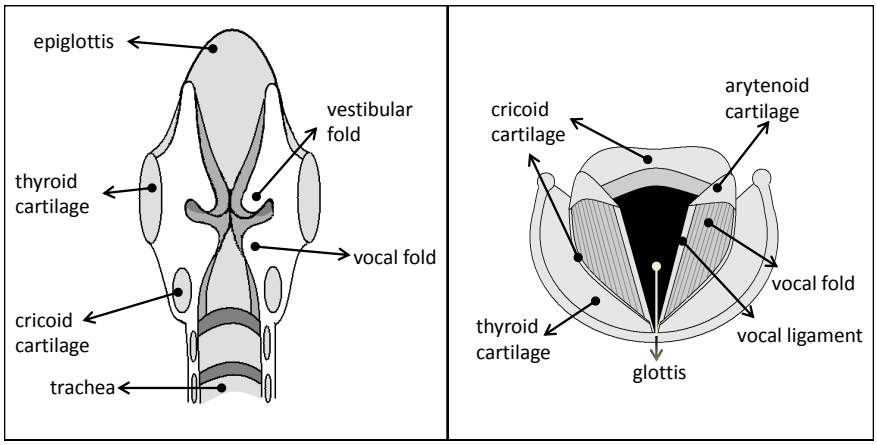

Figure 2.2: Vertical (left) and horizontal (right) transsection scheme of the larynx. After [25, 26].

subglottal and a supraglottal part and are the most essential organs in the phonation process. During respiration, the vocal folds are in abducted position (separated), leaving a gap between them which is called the glottis. During phonation, the vocal folds adduct (move together), narrowing the glottis. This makes the subglottic pressure to build up below the vocal folds. When the pressure is high enough, the vocal folds are forced to separate and the airstream is allowed to flow through the

vocal folds. The airstream through the vocal folds then accelerates causing a drop in pressure according to Bernoulli's theory. This drop in pressure then forces the vocal folds back together [25]. Subglottic pressure then builds up again and this cycle repeats at a certain frequency, called the fundamental frequency, until the vocal folds are relaxed or when there is no airflow anymore. The movement of the vocal folds during the glottal cycle is a very complex 3-dimensional movement, and different parts of the folds move in different parts of the cycle, creating a rich spectrum with harmonics.

2.2.1 Changing fundamental frequency and intensity

The fundamental frequency (F_0) of the voice can be altered by the muscles surrounding the vocal chords. These muscles can change their length, thickness and thus tension. The intensity of the voice can be regulated by three mechanisms [27]: controlling the subglottal pressure, changing the portion of the phonation cycle during which the glottis is open and altering the shape of the larynx. The first two mechanisms both have an effect on the behavior of the vocal folds during phonation, inducing an increased intensity of the voice which is often linked to an increase in F_0. The third mechanism consists of changing the shape and thus the formant settings of the vocal tract so that its resonance frequencies (the formants) coincide with the harmonics of the fundamental frequency. According to [28], fundamental frequency in male speech ranges between 90 and 165 Hz, in female speech between 158 and 259 Hz.

2.2.2 Unvoiced sounds and whispering

For the production of unvoiced sounds, like e.g. /p/, /f/ etc., the vocal folds are - like during respiration - in abducted position. While the expelled air passes through the larynx, the vocal folds do not vibrate and there is no phonation. The process is in reality much more complicated than this. Larynx and pharynx (throat) position are different compared to the production of voiced sounds, oral air pressure and air flow are higher/faster but the main audible difference is the lack of phonation [29].

A different usage of the glottis can be noticed during whispering. The vocal folds are then partly adducted and partly abducted. This leads to the generation of the very turbulent air flow characterizing /h/ and whispered sound.

2.3 Resonance

The expelled air vibrations continue their way out through the so-called vocal tract, which literally means the pathway of the voice. It consists of the laryngeal cavity, the pharynx, the oral cavity and the nasal cavity (see Figure 2.3). Different parts

of the vocal tract have different resonance frequencies, which will form the timbre of the air vibrations created by the vocal folds. Some parts of the vocal tract, like the nasal cavity and the nasopharynx (between nasal and pharyngeal cavity), can not be moved voluntarily but they can resonate according to their rather fixed dimensions. Other parts of the vocal tract can strongly adapt their dimensions, like the sinuses, the upper part of the larynx, the pharynx, and the total oral cavity to the desired sound. By moving the larynx, epiglottis (top of the larynx), velum, tongue, mouth musculature, jaws and lips during articulation, the cavities can considerably change shape, changing the resonance frequencies as well [25]. The soft palate

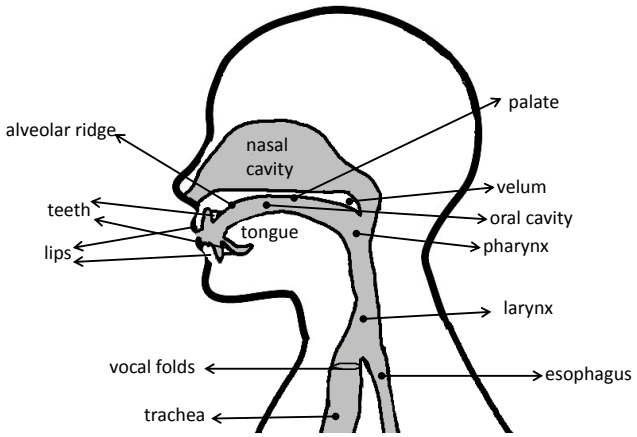

Figure 2.3: Vocal tract. The grey area is filled with air. After [11].

or velum plays a very important role in resonance. This soft tissue in the back roof of the mouth determines whether oral and nasal cavity should be connected or not. For the production of so-called non-nasal sounds (see Section 2.4), the velum retracts and elevates to separate the oral cavity from the nasal cavity in order to produce the oral speech sounds. For nasal sounds, such as nasal vowels /Eɶ/, /Oɶ/, /Aɶ/ and the consonants /m/, /n/ and /N/ [25], the velum does not close off completely, leading to an air flow and resonance through the nose, causing speech to be perceived as nasal.

2.4 Articulation

After this last phase of the speech production process, the expelled air vibrations are transformed into specific phones. Jaws, lips, tongue and other oral muscles/tissues (like the palate) are configured in a certain way to shape the oral, pharyngeal and nasal cavities according to the targeted phone. The outcoming sound will depend on the shape of the cavities and on how much and where they are narrowed.

The degree of narrowing (stricture) and the way the air flows out describes the *manner of articulation*, while the place of stricture determines the *place of articulation*.

2.4.1 Manner of articulation

As mentioned above, the manner of articulation describes the degree of constriction of the oral and/or pharyngeal cavity. Several categories can be distinguished [29]:

Plosives or stops are characterized by a short stop of the airflow during which the air pressure at some place in the vocal tract is built up, immediately followed by a short explosion during which the air is released. This process intrinsically consists of two phones, namely a closure followed by a burst. Examples of stops are /p/, /b/,/k/, /t/, /d/.

Fricatives are sounds which are characterized by a turbulent air stream caused by the air flowing through a small gap. Examples are /f/, /v/, /s/, /z/, /S/, /Z/.

Liquids are sounds for which the air flow is only slightly constricted. Depending on the way the constriction is made, this category is subdivided in **laterals** and **trills**. A lateral is a sound with an occlusion somewhere in the axis of the tongue but in which the air can flow along the sides of the tongue. An example of this is /l/. A trill sound is produced by vibrations between the tongue and the place of articulation. Examples of this are the Flemish "tongue-tip" alveolar /r/ and the French uvular /R/.

Taps and flaps are produced by briefly brushing the tongue towards the alveolar ridge or palate horizontally (flap) or vertically (tap). Often both terms are used as synonyms. These phenomena do not occur in Flemish. In the North-American pronunciation of e.g. /latter/ the middle /t/ can be replaced by a flap.

Nasals are produced with an occlusion in the mouth while the air is (partly) escaping via the nose because the velum is lowered. Examples of nasals are /m/,/n/ and /N/.

Approximants are produced with an even smaller, hardly noticeable occlusion. Their sound approximates the sound of vowels, e.g. /j/ sounds like /i/, /w/ like /u/.

Vowels can be distinguished from all categories mentioned above - all consonants - because they are realized without any obstacle for the air flow. They will be described in Section 2.4.3.

2.4.2 Consonant place of articulation

For consonants, the place of constriction in the vocal tract determines the place of articulation. Consonants can be produced at the following places [29], also indicated in figure 2.3:

Bilabial sounds are produced by letting both lips make contact, like in /m/, /p/, /b/ and the Flemish /w/.

Labiodental sounds are produced by using the lower lip and teeth, like in /f/ and /v/.

Dental sounds are produced by placing the tongue tip against the teeth, like in the English /th/. No Flemish phones are produced dentally.

Alveolar sounds are produced when the tongue tip hits the alveolar ridge, which lays right between the upper teeth and the palate. Examples are /t/, /s/, /n/, /d/, /l/, /z/, /r/.

Post-alveolar sounds are produced when the tongue hits both the alveolar ridge and the front of the hard palate. An example of this is /S/.

Retroflex sounds are produced by curling the tongue backwards behind the alveolar ridge. The English /r/ is formed this way. No Flemish phone is retroflex.

Palatal sounds are produced by pushing the tongue body against the palate, like in /j/.

Velar sounds are produced pushing the tongue towards the soft palate (velum), like in /N/,/k/,/g/,/x/ and /G/.

Glottal articulation is found in /h/ and the glottal stop (like just before an initial vowel or between two identical vowels). In these cases, the glottis is closed or narrowed for the articulation.

The IPA chart for consonants (Figure 2.4) gives an overview of all existing consonants for all possible languages, classified according to their manner and place of articulation. Voicing is also indicated.

THE INTERNATIONAL PHONETIC ALPHABET (revised to 2005)

CONSONANTS (PULMONIC) © 2005 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

Figure 2.4: IPA chart for consonants.

2.4.3 Vowel articulation dimensions

In vowel production, the air can flow without much obstruction through the vocal tract. The configuration of the nasal, oral and pharyngeal cavity will determine the resonance frequencies and thus the produced vowel. The shape of the cavities is mainly determined by the amount and place of elevation of the tongue (leading to vowel height and place), the rounding of the lips and the opening of the velum.

Vowel height. Elevation of the tongue leads to a separation between the oral and pharyngeal cavity. The air flows from the pharyngeal to the oral cavity through a narrower opening, demarcated by the tongue body and the roof of the mouth. The vertical position of the tongue determines how narrow the opening is. This in turn determines the so-called vowel height. High vowels, also called closed vowels, are produced with a high position of the tongue, leading to a close, narrow opening. /i/ is an example of this. Low or open vowels are produced with a low position of the tongue, leading to a wide opening between pharyngeal and oral cavity. /a/ is an example of this. Vowel height is a continuous dimension. All positions between low and high are also possible, all leading to other vowels. Often, four vowel heights are distinguished: high, mid-high, mid-low and low (or close, close-mid, open-mid and open). However, this distinction is rather artificial and is just made to roughly classify vowels.

Vowel place. The place of the tongue body along the horizontal dimension determines the so-called vowel place. /i/ is produced in the front of the mouth, while /u/ is pronounced in the back of the mouth. Again a continuum of vowel places is possible, but often one distinguishes front, mid and back as possible places.

Rounding. Vowels can be pronounced with rounded or spread lips. /i/ and /y/ have the same vowel height and place, but the first is spread and the second one rounded.

Nasality. If the velum is opened, part of the air escapes through the nose, leading to nasal vowels. Generally, vowels are not nasalized in Flemish. However, there are quite some French words used in Flemish which are nasal, like /parfum/, /bulletin/, etc. using /Y$\tilde{}$/, /E$\tilde{}$/ etc.

Diphtongs. Diphtongs are sounds which gradually shift from one vowel to another. Three diphtongs exist in Flemish: /E+/, /Y+/ and /A+/.

The IPA chart for vowels (Figure 2.5), gives an overview of all existing vowels

for all possible languages, classified according to vowel height and place. Rounding is also indicated.

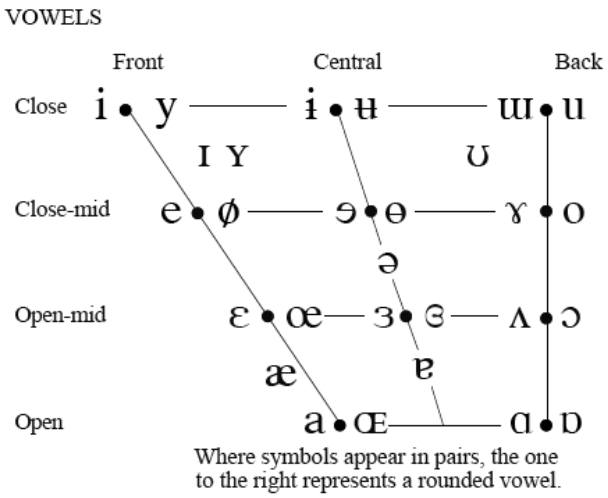

Figure 2.5: IPA chart for vowels.

3
Speech Pathologies

Chapter 2 makes clear that the mechanism of producing speech is dependent on the well-functioning of a whole series of nerves, muscles etc. of the speech organs. One little bug in this process can disturb the speech production, leading to a speech disorder. In this chapter, some of the speech disorders I was confronted with during my research will be described: dysarthria, laryngectomy, hearing impairment and cleft lip and/or palate. For every pathology, I will start with a general description before discussing how the four speech production sub-processes described in the previous chapter are affected.

3.1 Dysarthria

Dysarthria is a so-called motor speech disorder, meaning that the origin of the disorder can be found in the motor part of the nervous system. It causes a malfunctioning of one or more muscles used to produce speech. Note that the problem is situated in the speech production, not in the speech planning.
Dysarthria can be caused in many ways. A stroke or trauma can damage part of the motor nervous system, a biochemical disorder can change the production of neurotransmitters, a virus or tumor can damage the nervous system or it can be congenital. The exact symptoms, the change in force, initiation, speed, coordination and tonus of muscle movements depend on the exact location of the "damaged" region.

3.1.1 Flaccid dysarthria

When one or more of the cranial nerves in the peripheral motor speech system are damaged, all input of the central nervous system which needs to be transmitted through that/those nerve(s) is hindered or blocked. This is reflected in all automatic and planned movements of the muscles innervated by the part of the nerve after the lesion. Those muscles will be partially or totally paralyzed, weak, hypotone (low muscle tone or tension), sometimes with small fasciculations (involuntary contractions and relaxations) and degenerating over time. Symptoms depend on which nerves are damaged and how severe the damage is. Most common symptoms in dysarthric speech production are listed in Table 3.1. Note that "phonation" refers to the quality of the sound produced by the vocal folds, while "voicing" is considered as part of the articulation process since the ability of making a clear distinction between voiced and unvoiced sounds is partially related to the articulatory movements of the mouth.

3.1.2 Spastic dysarthria

Spastic dysarthria results from damage to the voluntary part of the central motor system, namely the motor cortex and/or the pyramidal system. Since both of those are organized in an contralateral body scheme, the damaged region will cause a problem in the contralateral muscles. As the peripheral nervous system receives a bilateral innervation for every movement, the problem will be mild if the lesion is unilateral. Otherwise, it will be severe. The principle result of the damage is difficulty with all fine motor movements. As the motor cortex has an inhibiting influence on the muscle tone, a lesion in this region will lead to exaggerated stretch reflexes, resulting in increased muscle tone and decreased coordination, hence the name spastic dysarthria. In general, the disorder affects/paralyzes the facial muscles, leading to a non-expressive, drooling face, with difficulties with swallowing and moving the lips and the velum. Patients sometimes laugh or cry unprovokedly, which is caused by the disturbed inhibiting influence of the motor cortex. Most common symptoms in speech production are listed in Table 3.1.

3.1.3 Ataxic dysarthria

The cerebellum is responsible for monitoring the precision, smoothing and coordination of movements. A lesion in this area will thus disturb coordination, affecting force, speed, range, timing and direction of the movements [22]. The resulting pathological speech is known as ataxic dysarthria. Typical speech symptoms are listed in Table 3.1.

3.1.4 Hypokinetic dysarthria

A part of the extrapyramidal system produces dopamine, an important neurotransmitter. In case of an underproduction of dopamine, like e.g. in a patient with Parkinson's Disease, the inhibition of muscle tone and the initiation of movements is weakened. This causes a higher muscle tone (hypertonicity) and a slower initiation of movements, hence the name hypokinetic. All muscles are hypertonic, leading to rigid muscles. As the extrapyramidal system regulates automatic, involuntary movements, mostly automatic movements will be slowed down, weakened, limited. Another typical symptom is the tremor (at about 6 Hz). In speech production, the most striking symptoms are listed in Table 3.1.

3.1.5 Hyperkinetic dysarthria

Both hypo- and hyperkinetic dysarthria originate in the extrapyramidal system. While the first is due to an underproduction of dopamine, slowing the initiation of movements down, the latter is due to a lesion in this area disturbing the inhibition of movements in the opposite way, causing more movement/activity. This leads to hyperkinetic symptoms. Speech production symptoms are listed in Table 3.1.

3.1.6 Mixed dysarthria

If multiple regions in the nervous system are affected, one can develop a mix of the dysarthrias described above. Speech production will be affected according to the mix of dysarthrias.

3.2 Laryngectomy

While dysarthria points to a speech disorder, laryngectomy refers to a surgery leading to speech problems. This surgery is almost always a result of laryngeal cancer and involves complete or partial removal of the larynx. While a total laryngectomy used to be performed even for small cancers, the tendency nowadays is to save the "voice box" as much as possible and to use radiation and chemotherapy for smaller cancers [30]. As in dysarthria, the name "laryngectomy" covers a wide range and variance of pathological speech symptoms, this time not only depending on the severity of the cancer and thus of the surgery, but also on the used substitute speech after surgery. In the next sections, I will describe both partial and total laryngectomy and the consequences for speech after surgery.

type	localization	phonation	resonance	articulation
flaccid	peripheral motor system: cranial nerves	breathy and harsh, monotone, short sentences, low volume	hypernasality	weak articulation, imprecise consonants, insufficient lip closure, reduced tongue force
spastic	motor cortex or pyramidal system	breathy, harsh, strained, strangled due to higher muscle tone, monotone, low pitch, bursts of loudness, slow rate	hypernasality	imprecise consonants, malformed vowels, reduced tongue strength, reduced length of voiceless segment in voiced plosives
ataxic	cerebellum	harsh voice, variations in loudness and pitch, monotone, irregular breathing, longer pauses, every syllable is stressed	normal	imprecise consonants, malformed vowels, prolonged phonemes, varying goodness of articulation
hypokinetic	extrapyramidal system	breathy, hoarse, low volume, low pitch, monotone, accelerating speech, sometimes in short rushes	hypernasality	imprecise consonants, starting and stopping of speech can be troubled
hyperkinetic	extrapyramidal system	breathy, harsh, strained, strangled, voice stoppages, monotone, too much or too little stress, inappropriate voicing and pausing	hypernasality	imprecise consonants, malformed vowels, varying goodness of articulation
mixed	2 or more of above	mix of above		

Table 3.1: Dysarthria subtypes and possible symptoms. Based on [20, 22].

3.2.1 Causes

As described before, the larynx can be divided into a subglottal, glottal and supraglottal area. Laryngeal cancer can originate in all three regions, but glottal cancer occurs the most [26]. Four stages of severity can be distinguished:

- T0: only a superficial layer is affected (called carcinoma in situ).
- T1: only the glottal area is affected, one or both vocal folds are affected
- T2: The supra- or subglottal area is affected as well, the vocal folds can be less mobile
- T3: one or both vocal folds are fixated (immobile)
- T4: the tumor has metastases outside the larynx

Nowadays, small tumors (stage T1 or T2 or even some T3) will be fought with radiation therapy. If this does not work, a partial laryngectomy will be performed. Tumors of stage T3 and T4 will often lead to a total laryngectomy.

3.2.2 Partial laryngectomy

In this surgery, only a part of the larynx is removed to save the voice box and swallowing functions as much as possible to keep life quality as good as possible. There are different types of partial laryngectomy procedures, depending on the type, location and metastases of the cancer.

CO_2 laser surgery can be used to treat some stage T0 to T2 cancers. An endoscope is passed down the throat to locate the tumor, which is then excised using a high-intensity laser on the tip of the endoscope [30, 31]. Although laser surgery is an organ-preservation method, research indicates that voice quality after this kind of partial laryngectomy is not better than after total laryngectomy [32]. It shows that some parameters like shimmer and jitter are worse or equally bad as for total laryngectomees with tracheo-esophageal speech (See Section 3.2.3). Intelligibility however is considerably better after laser surgery [32].

Horizontal or supraglottic partial laryngectomy is an operation in the horizontal plane to remove part of the supraglottis: the epiglottis (top of the larynx), false vocal cords, and superior half of the thyroid cartilage. Vocal cords are saved. Swallowing and aspiration problems are common after this operation, but voicing and articulation should not be affected [31].

Supracricoid laryngectomy includes removal of the entire supraglottis, the false and true vocal cords, and the thyroid cartilage. Maximum one arytenoid (musculature around the vocal cord which regulates the frequency and position of the vocal cord) may be resected. The remaining arytenoid(s) is/are sutured against the tongue base, which will now serve as phonation organ instead of the resected vocal cords. Respiratory function is dependent on the preservation of the cricoid cartilage [31]. Although supracricoid laryngectomy is preferable over a total laryngectomy as it does not need a permanent tracheostoma (see 3.2.3) and main functions such as breathing, swallowing and phonation are (mostly) maintained, speech performance after this surgery is very variable and not better and sometimes even worse than after a total laryngectomy [33–35]. While the fundamental frequency is lower than in laryngeal speech, jitter and shimmer are much higher and intensity and phonation time are very small [35].

Vertical partial laryngectomy is a vertical resection of the larynx. Usually, the thyroid cartilage is divided medially and one false vocal fold and one true vocal fold are removed (hemilaryngectomy). Sometimes the vertical plane has a different orientation, removing the anterior 1/3 of the folds and cartilage (frontolateral laryngectomy). Unlike all other types of laryngectomy, fundamental frequency after this type of surgery is on average higher than in laryngeal voice. This is due to thinner, lighter and/or shorter vocal folds [36]. Jitter and shimmer are very elevated again, sometimes even worse than after a total laryngectomy [37]. The voice is also likely to be dysphonic but articulation should be unaffected [31].

3.2.3 Total laryngectomy

In this surgery, the complete larynx is removed. This includes all cartilages, folds, hyoid bone, epiglottis and tracheal rings. An opening or stoma is made in the trachea in the front of the neck, separating upper and lower airways permanently. Air enters and leaves the trachea and lungs through the stoma. As the voice box is completely removed, the patient will need to use another source for phonation. Moreover, the complete separation of pharynx and trachea (see figure 3.1) also omits the air passage through the oral cavities where resonance and phonation take place. The patient will thus have to find an alternative way to speak, such as esophageal speech, tracheo-esophageal speech, or to use an electrolarynx. All three mechanisms will be described in the next paragraphs.

Esophageal speech The cheapest way of phonation after a total laryngectomy is esophageal speech (E-speech). In this method, air is injected into the esophagus and then expelled through the mouth. While the air passes the top of the esophagus, which functions as the so-called pseudo-glottis, the local muscles will contract

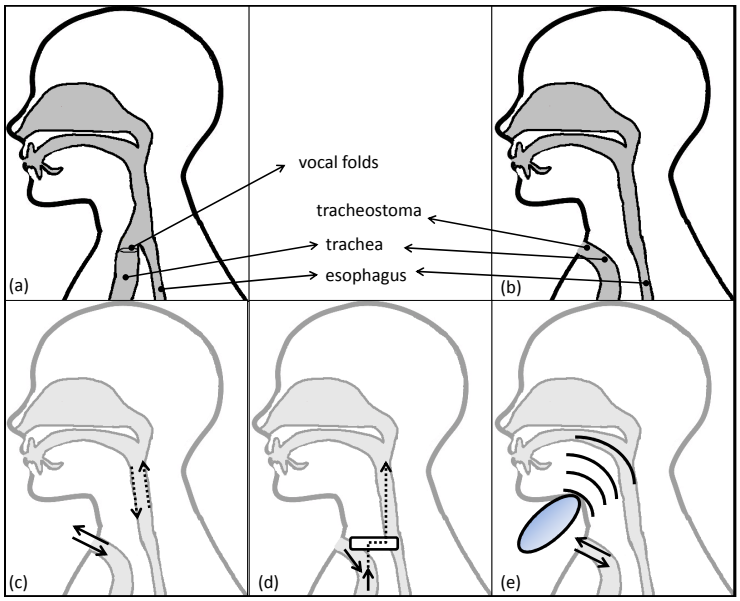

Figure 3.1: Total laryngectomy and substitute speech possibilities. (a) Scheme of the head-neck area before laryngectomy. (b) Scheme of the head-neck area after laryngectomy: trachea ends in a stoma and is separated from the oral and pharyngeal cavities. (c) Esophageal speech. (d) Tracheo-esophageal speech (e) Electrolarynx. Full arrows denote in- and exhalation, dotted arrows denote air injection and expulsion for speech purposes.

causing vibrations to produce phonation. While this technique is the cheapest as no battery is needed (as in electrolarynx) and speech is hands-free (unlike tracheo-esophageal speech), it is not the easiest one as intense speech therapy is needed to learn how to force the air into the esophagus. Three methods are used for this [38]:

- Injection is a method where air in the mouth is compressed by backward and downward movement of the tongue to be injected into the esophagus.
- Swallowing air into the esophagus is a second method. Drawback of this method is that it takes more time to transport the air in and out again.
- Inhalation is a method that exploits the fact that the pressure in the esophagus is lower than in the trachea after inhaling air into the stoma. Because of this underpressure, air enters the esophagus when the pharyngo-esophageal (PE) muscle is relaxed during inhalation.

Success rates for E-speech vary between 14% and 75% depending on the details of the surgery and the health of the patient [39–41]. It requires 30 to 50 hours

of intense speech therapy to master the technique, and the therapy often fails. This can be due to a wide range of reasons, both physically (e.g. reflux) and mentally [38, 41].
Apart from the learning difficulties, a major drawback of esophageal speech is the small air volume per injection: the air volume per injection is only about 50 (for frequent injection) to 85 ml (for one injection only) while the capacity of the lungs can reach about 3 liters [38]. This strongly reduces the length of an utterance to maximum 2 seconds [42].
As the vocal source in esophageal speech (pseudo-glottis) is located in the top of the esophagus, the produced sound will be very different from "normal" laryngeal speech. Most striking characteristics are the following:
Phonation. First of all, the pseudo-glottis vibrates at a frequency of on average 60-80 Hz [38, 43], which is far below the normal pitch range of 90-165 Hz in healthy male speakers and 158-259 Hz in healthy female speakers [28]. Jitter and shimmer are also higher than in normal speech, while the amplitude and intonation of the voice are reduced [43].
Resonance. The velopharyngeal function is altered [44]. Because of the small air volume in the mouth, esophageal speakers produce nasal phonemes with less coupling of the oral and nasal cavities than nonlaryngecomtized speakers, which will lead to a reduced or missing nasal air release, also called hyponasalization [45]. The same mechanism causes a higher oral pressure and no nasal airflow during pronunciation of /p/.
Articulation. Apart from the reduced intelligibility of nasals [44], esophageal speakers face other articulatory problems as well. As they do not have vocal folds, they have difficulties in distinguishing voiced and voiceless consonants [46, 47]. Most errors occur in the voiceless stops, which are replaced by their voiced counterpart, e.g. /p/ and /t/ will often be confused with /b/ and /d/ respectively, while the voiced versus voiceless confusion is less frequent. Voiceless fricatives can also be confused by their voiced counterpart, but not as often as in stops [47]. According to [46], the voice onset time, defined as the time interval between the release of stop occlusion and the onset of the following vowel, is significantly shorter than in normal speech. This can be explained by the fact that the pseudo-glottis (PE - segment) is normally adducted when not involved in vibration while the vocal folds are abducted in rest. In addition, the air volume between the sound source and the lips is smaller than in laryngeal speakers. This means that the air pressure drop needed to end the burst is reached in less time and the vibration of the pseudo-glottis is started easier and earlier.

Electrolarynx The electrolarynx was introduced in the 1940s, at a time when tracheo-esophageal speech (TE-speech, See Section 3.2.3) had not yet been invented and esophageal speech was the only option in speech recovery. Since

E-speech is not always possible, the electrolarynx became a popular alternative. Nowadays, TE-speech is the most popular method, but for 11% [44] to 50% [48] of the laryngectomees the electrolarynx is still used, sometimes just in the first weeks after operation, sometimes as a backup method and sometimes as the only method. The electrolarynx is a hand-held device with a battery-driven vibrator, which is placed against the neck or into the mouth by a tube. Although the intra-oral type leads to a better speech quality, the neck-type is used more often because it is easier and more comfortable to use [40]. The vibrations are passed into the pharynx, replacing the vibrations of the vocal folds. The speaker then articulates with the tongue, palate, throat and lips as usual [49]. This technique is very easy to learn as the patient only has to learn where and how to place the device correctly. Another advantage is the unlimited utterance length as electrolaryngeal speech (EL-speech) does not need pulmonary support. Although EL-speech is very easy to acquire, it suffers from many problems. First of all, the device is not hands-free. Secondly and more important, it is known for its mechanical and monotonous sound, cannot control the pitch and produces noise. It is generally agreed that EL-speech is less intelligible than E- and TE-speech [40, 50]. In the last decade, some devices have been developed to solve the monotonicity problem. Electrolarynges with finger-tip pitch control are already commercially available. Electromyographic and expiration pitch control are still in the experimental phase [40]. Those last two use effects happening naturally in laryngeal speakers: neck muscle activity changes and expiration pressure raises when persons raise their voice. Watson et al. [48] found that EL-speech with a variable fundamental frequency was more understandable by the listeners because it sounded more natural and the pitch variations indicated the important words. This prosodic improvement does however not change intelligibility. Most important phonation, velopharyngeal and articulation characteristics of EL-speech are the following:

Phonation. The fundamental frequency of EL-speech depends on the type of device. It varies between 80 and 125 Hz and is monotonous. Although some new devices can adapt intonation, the frequency changes still do not meet the normal frequency changes [49]. The signal-to-noise ratios are also very low compared to normal speech.

Resonance. Information about resonance function in EL-speech could not be found.

Articulation. On the level of articulation, the main problem lies again in the voicing domain. Because the electrolarynx is continuously pulsing during an utterance, the voiceless sounds are perceived as being voiced. This is the most predominant in voiceless stops, confused with their voiced cognates [39]. Voiceless fricatives are less confused with their voiced counterpart. As EL-speech is independent from breathing, patients have to take extra care of the fricatives and have to learn how to use the air in their mouth to create strong frication noise without pulmonary

airflow. Fricatives are the second least intelligible (after stops) [39]. Place and manner errors are not so common in EL-speech.

Tracheo-esophageal speech Before 1980, the only valid options after a total laryngectomy were esophageal speech and the use of an electrolarynx (see Section 3.2.3). Since then, voice prostheses have become a valid alternative. Nowadays it is even the preferred option in many countries [42, 51]. A voice prosthesis consists of a one-way valve placed in a surgically created hole between the trachea and the esophagus. The prosthesis keeps food out of the trachea and lets air from the trachea into the esophagus for so-called tracheo-esophageal speech (TE-speech). As in esophageal speech, the pseudo-glottis is the PE-segment, but the main difference lies in the air source, being the lungs instead of the esophagus, making TE-speech more natural for the patient and also easier to learn. The patient can already start speaking a couple of days after his/her operation. As the lungs can contain a couple of liters of air, the utterances can be significantly longer than in E-speech (using the 60-80 ml of the air injected in the esophagus) [42]. Voice prostheses can be classified into two categories. *Non-indwelling* prostheses can be removed, cleaned and placed back by the patient himself/herself, while the patient needs a doctor to do so for an *indwelling* prosthesis. As this last type is easier to clean and maintain for the patient and is designed to be more robust, the indwelling prostheses are usually used nowadays [42]. Among the most popular prostheses are the Provox and Provox II valve, the Groningen prosthesis and the Blom-Singer prosthesis, all made to decrease the discomfort for the patient as much as possible by optimizing the speech and voice (low airflow resistance) characteristics and improving lifetime and hygiene of the device [52, 53]. Currently, the average lifetimes of prostheses vary between 5 to 6 months. This relative short life is caused by candida growth on the silicone parts of the prosthesis. Because of this, the prosthesis does not fully shut the trachea off anymore, allowing leakage of fluid to the trachea, making the patient cough. Anticandida and yogurt and other probiotics could slow down the growth of candida [42, 54]. Another discomforting aspect of the prostheses is the closing. Usually TE-speakers have to close the stoma with a finger directly to divert the expiratory air through the PE-segment. Nowadays, an automatic valve makes it possible to talk "hands-free". This valve closes the stoma whenever the air pressure rises (which indicates that the patient is trying to speak) [42].
TE-speech is claimed to be the best intelligible alaryngeal speech [51,52]. Still the absence of vocal folds and usage of the PE-segment as pseudo-glottis introduces some typical articulatory and phonatory problems:
Phonation. As in E-speech, the fundamental frequency of speech is on average 80- 100 Hz, which is lower than in laryngeal speech. Jitter and shimmer are higher than in normal speech, but already lower than in E-speech.

Resonance. While coupling of oral and nasal cavities is reduced in E-speech, [45] indicates that the nasalization of TE-speech is closer to normal. This could be due to the fact that TE-speakers dispose of a higher air volume per utterance than E-speakers.
Articulation. The main articulation problem lies again mainly in the voiced-voiceless distinction. TE-speakers generally realize manner and place of articulation of plosives and fricatives the right way, but voiced consonants can be perceived as devoiced and - more often - voiceless consonants are perceived as their voiced counterparts [51, 55]. This could be explained by the tendency of the PE-segment to vibrate [51]. Jongmans [55] also found other kinds of confusions for nasals and approximants, noticing especially a high percentage of missing fricatives, nasals and approximants in word final positions of a word-based test. This could be because the intensity drops and articulation is less precise at the end of an utterance.

3.3 Cleft lip and palate

Cleft lip and cleft palate or a combination of both are quite common facial malformations during embryonal development. They consist of a cleavage of lip, jaw, palate or a combination of those. The cleft can be uni- or bilateral and exists in many severities, ranging from a small dip in the lips to a cleavage from nose to uvula [56]. Figure 3.2 shows several types of cleft. A cleft palate makes the separation of oral and nasal cavities impossible. Incomplete separation will lead to disturbed swallowing, sneezing, breathing and blowing/whistling [56]. Children born with cleft lip and/or palate will be monitored throughout their childhood by a team of surgeons, speech therapists and social workers to make its development as normal as possible despite all these problems. They will have to undergo a number of surgeries to close the lip, soft and hard palate and possibly correct jaw and teeth structure. This whole process is spread over the whole childhood, which causes retardations and alterations in speech and language development [57]. As described in 2.3, the separation is needed for proper nasalization of sounds. If the velum is not capable of closing off the nasal cavity from the oral cavity, all speech will sound hypernasal, as nasal resonance will occur more and stronger than intended. Speech of patients with an insufficient velum closure have some typical speech characteristics:

Phonation. During consonant production, the oral airflow is often accompanied by or even replaced by a (sometimes audible) nasal airflow. It is best perceived in voiceless stops and fricatives which require some air pressure, which then leaks away through the nose. In voiced stops and fricatives, nasal turbulence can be observed [58, 59].
Resonance. Hypernasality is considered the primary characteristic of cleft palate speech, as described above.

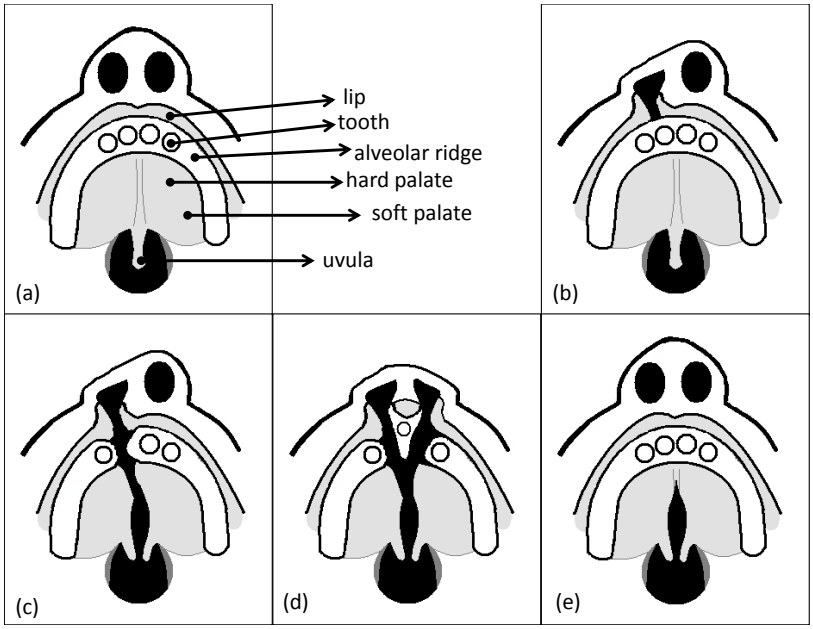

Figure 3.2: Different types of cleft. (a) Normal lip and palate. (b) Unilateral cleft lip. (c) Unilateral cleft lip and palate. (d) Bilateral cleft lip and palate. (e) Cleft palate.

Articulation. Articulation errors are manifold in cleft palate speech [56, 58]:

- First of all, articulation is weak due to the reduced intraoral pressure. This is perceived best in voiceless stops and fricatives.
- Double articulation occurs when a single contact between tongue and palate is not enough to realize a consonant. The patient will then make two points of contact, leading to a double articulation, e.g. /tk/,/dg/ etc.
- Another common error is the backing of consonants which are normally produced at the front. /t/ can then be produced as /k/ and /k/ can be produced at the uvular or glottal position. This backing process is typical for cleft palate children and could be due to abnormal neuromotor learning due to structural abnormality [58].
- Another consequence of reduced intraoral pressure is compensatory articulation. Only glottal and pharyngeal plosives and fricatives are not produced orally and can be realized correctly. Uvular and velar fricatives shift to more glottal or pharyngeal positions, /s/ is frequently replaced by a palatal or lateral fricative, and /f/, /s/, /S/ and /x/ are sometimes produced as a voiceless nasal fricative (no or little oral airflow).

- Nasal substitution is yet another consequence of the pressure reduction. Oral sounds are replaced by nasal sounds, e.g. /b/ becomes /m/, /d/ becomes /n/ etc.

- Sometimes consonants can be lateralized or palatalized. In those cases, the tongue is in a lateral position between the teeth or touching the palate during articulation, like e.g. /s/ becomes /S/ etc. This is considered as a minor articulation error.

As mentioned before, cleft palate and lip can be corrected by surgeries like palatoplasty and velopharyngeal flap surgery [56]. While all surgical methods show good results in the recovery of the clefting, the effect on the speech outcome can vary, maybe also depending on the type of the surgery [60].

3.4 Hearing impairment

A last pathology I will briefly describe, is hearing impairment. Hearing disorders are classified along three dimensions: the origin of the disorder, the severity and the age and gradient of onset.
According to the origin of hearing disorders, two major types can be distinguished: hearing sensitivity loss and suprathreshold hearing disorder [61]. Most common is the hearing sensitivity loss, meaning that the ear is less sensitive in detecting sound. Depending on the exact location, hearing sensitivity loss can be conductive or sensorineural. The first type refers to an attenuation of sound in its mechanical form as it travels through the outer ear. The second type refers to a problem in the inner ear during transduction from mechanical pulses to neural impulses, the form in which the hearing passes its data to the hearing nerve to the brain. Causes of hearing sensitivity loss are manifold: both (genetic) embryologic malformations as structural changes secondary to infection or (acoustic) trauma are possible causes, as are many others [61]. Some toxins and medications can also be a cause. Suprathreshold hearing losses result from lesions in the nervous system affecting the hearing nervous system. They are caused by a tumor, stroke, trauma or a developmental disorder or just by aging.
Hearing impairment is quantified by the use of an audiometer. This measures the hearing sensitivity across the range of audible frequencies and its deviation from normal. It returns a degree of loss, ranging from mild (24-40 dB hearing loss) to profound (more than 90 dB hearing loss). More specific information such as the curve of the spectrum of hearing loss, are also returned. This information enables the speech therapist to predict which phones will be heard and which will not. It also determines whether the hearing loss is unilateral or bilateral - in one or both ears.

Crucial for the impact of hearing loss is the age of onset. If it occurs before linguistic development, also called prelingually, and if the hearing loss is profound enough and intervention (e.g. cochlear implant, see further) is not done early enough, the prognosis for adequate language and speech development is diminished. If the hearing loss occurs after linguistic development, called postlingually, the prognosis is much better. Apart from the age of onset, the speed of onset is also very important [61, 62]. The slower the hearing loss develops, the more time the patient has to develop compensatory strategies for hearing (like lipreading and learning contrasts between phones in the new situation) as well as for speaking.
Hearing losses can be (partially) compensated by the use of aids to hearing. Conventional hearing aids basically consist of a microphone, an amplifier and a loudspeaker, sending the incoming signals in an amplified version to the ear. They are most often worn behind the ear. If the deafness is more profound, the problem is mostly situated in the transduction of the sound waves into electrical impulses. In this case, a cochlear implant might work. This consists of an external microphone, amplifier and transmitter sending the electrical impulses to a receiver or electrode which has been implanted into the cochlea to artificially stimulate the hearing nerve. [61].

Speech characteristics of hearing impaired speakers Severity as well as age and gradient of onset are known to affect speech intelligibility of the hearing impaired speakers as described above. Typical speech errors for hearing impaired speakers are manifold and involve phonation, resonance and articulation.
Phonation. The fundamental frequency is often quite high in females, and slightly too low in male speakers. Moreover, some speakers show highly irregular, inappropriate and strong pitch variations up to 100 Hz within an utterance [62]. Sometimes breathiness is observed. This is due to an excessive airflow during voicing and slow vocal fold cycle, resulting in a turbulent waveform with high energy in the low frequencies and low energy in the high frequencies.
Resonance. Velopharyngeal control is known to be difficult for hearing-impaired speakers. Both hyper- as hyponasalization can occur [63], although excessive nasality is more investigated. Studies report as much as 76% of the profound hearing-impaired children producing excessive nasalization in at least half of the vowels, and clusters containing a nasal and a stop seem to be the most difficult [62].
Articulation. Hearing impaired persons tend to make many articulation errors. In general, the problem lies within the visibility of articulatory gestures. As the hearing impaired do not have auditory feedback to learn the right pronunciation, they need the visible aspect to understand how to produce a phone [62]. This explains why phonemes produced in the front of the mouth are more often produced correctly than phonemes located in the back of the mouth. The latter are then replaced by a phoneme with a close-by place of articulation. While place of articulation is

still in the right range, especially for the visible phonemes, manner of articulation is more problematic as coordination and timing of the movement of articulation is sometimes difficult. As mentioned above, errors in nasality are common: non-nasal phonemes are nasalized while nasal consonants are often replaced by a stop. Again, most errors occur in phonemes produced in the middle or the back of the mouth, like e.g. palatal plosives and fricatives and velar sounds. Alveolar sounds are also prone to errors. This is due to the fact that more sounds are pronounced centrally, which means that a higher accuracy is needed to produce those sounds correctly. Hearing impaired speakers often lack accuracy as they only have the visible aspect, explaining why phonemes produced in the middle of the mouth are also prone to more errors. Those phonemes are then substituted by another phoneme, or - more commonly in severely and profoundly hearing impaired - simply omitted.

One of the most frequent errors in consonant production is the voiced-voiceless confusion, most often voiced for voiceless substitution [62]. One hypothesis is that this is due to a failure in coordination of timing of respiration, phonation and articulation. More precisely, the voice onset time (VOT), being the time needed to start voicing after a stop, is in many cases equally short for voiced and for unvoiced stops, while the VOT should be longer after a voiceless stop. This leads to misperceptions of unvoiced sounds as being voiced.

Although vowel errors are less common than consonant errors, it is known that the formant frequencies of deaf children's vowels tend toward the central vowel /@/. The reduced phonological space is most present in the vowel place dimension. Front-back tongue movement is less visible than jaw movements for vowel height and thus less simple to mimic [62].

4

Intelligibility Assessment

In clinical practice there is a great demand for fast and reliable methods for assessing the communication efficiency of a person with a speech disorder. It is argued in several studies (e.g. [64]) that intelligibility is an important criterion in this assessment.

The concept of intelligibility is rather intuitive and leaves room for interpretation. The process of being understood depends on so many factors that covering them all in one single measure is impossible. This is reflected in the large number of tests and scales available for assessment of intelligibility, each measuring some contributing aspects which are then converted to a test-specific score giving an impression of the intelligibility. The test outcome - an intelligibility score - thus has to be interpreted within the framework of the used test, and cannot be considered as generally valid as another is bound to yield a different score.

Apart from the variety in available tests and scales, intelligibility can be measured on several linguistic levels, ranging from phoneme intelligibility to running speech intelligibility (RSI). While the first aims to measure the intelligibility of every single phoneme of a speaker, RSI measures the performance of the speaker while reading more natural speech material like sentences or paragraphs.

In Section 4.1, I will first discuss some key factors for differentiating between tests as well as some prerequisites for a reliable (perceptual) intelligibility test. I will then introduce the DIA, the Flemish test my work was based on. A number of automated intelligibility tests will be discussed in Section 4.2.

4.1 Perceptual Evaluation

Nowadays, clinicians mainly rely on perceptual intelligibility tests. As these tests are subjective in nature, all kinds of methods are used to make them as reliable as possible. One of the primary prerequisites for getting reliable scores is that the test is designed in such a way that the listener cannot guess the correct answer based solely on contextual information. That is why these tests use random word lists, varying lists at different trials, real words as well as pseudo-words, etc.

4.1.1 Variables in intelligibility testing

In this section, I will discuss some important parameters determining the linguistic level, reliability, rating and profoundness of an intelligibility test.

4.1.1.1 Level of intelligibility

Intelligibility can be assessed at different linguistic levels, ranging from phoneme intelligibility (PI) to running speech intelligibility (RSI). While the first investigates a person's ability to correctly pronounce every phoneme, the latter creates a more global view of a person's ability to communicate in a more realistic and natural setting such as running speech as it is used for daily communication. As one would tend to opt for RSI measures because of their naturalness, others argue that PI measures are preferable in case of severely impaired speakers who are still able to make themselves clear by uttering single words without having the strength to produce intelligible phrases. Clearly pronounced phonemes are therefore more important for them, making PI tests more preferable. PI tests also have the advantage of permitting to extract more detailed information about the underlying articulation problems. A conclusion is that each linguistic level of intelligibility testing has its own merits and contributes in its own way to the global assessment of the speaker [2].

Note that intelligibility of a person can vary with the linguistic level. Moderately pathological speakers can be fairly intelligible at sentence level but hardly intelligible at a syllable or word level. On the other hand, severely pathological speakers can score higher on single words than on sentences. This is reflected in an only moderate correlation between PI and RSI [2].

4.1.1.2 Predictability of test material and utterance

The level of intelligibility testing and - more general - the type of test material does not only influence the speaker's ease and quality of uttering the speech but also the listener's ability to understand it.

First of all, test sets consisting of existing words and/or meaningful sentences are to some extent predictable, which creates a positive bias in the intelligibility

score as the listener is able to guess the words or sentences by linguistic and contextual information. This can be solved by using sets of phonetically similar words or minimal pairs which are equally likely in the given context. Another possibility is to include non-existing words or sentences without any meaning.

Secondly, every kind of test material gets predictable for a speech therapist who is employing the same test regularly. The use of randomly generated test items can circumvent this problem. Both the speech therapist and the patient will then more often encounter new test items, reducing their familiarity with the test. As a consequence, randomized tests are more reliable and valuable [12].

A third concern is the presentation of the test items. Most commonly, the test items are presented in written form to the patient. This requires good reading skills, which reduces the applicability on children. Young children therefore often perform picture naming tests. A last option is repetition of the orally presented test items: the speech therapist says the test item and the patient repeats it. In the last case, the patient has obviously to be judged by a second person who did not hear the speech therapist's examples.

A last important factor is the examiners familiarity with the patient or with the type of pathological speech. This possible influence can again be decreased if the examinator is not the patient's therapist.

4.1.1.3 Rating of intelligibility

Concerning the rating of intelligibility, there are two possible strategies: scaling and measuring.

Scaling. The degree of intelligibility can be expressed on a Likert-scale [65]. Such a scale is often applied in ratings of RSI where it provides a quick, rough overall index of intelligibility (e.g. on a scale from 1 to 5 where 1 denotes totally unintelligible and 5 denotes very intelligible/normal). However, it does not provide any information on the underlying problem and it is often not precise enough to monitor progress during speech therapy [66].

Measuring. Another approach in intelligibility assessment is to measure and quantify the degree of intelligibility by means of a quantitative score, e.g. a percentage of correctly recognized items. Such a score provides a more accurate index of the functional limitation of the speaker. It is mostly used for word or phoneme intelligibility assessment. Measuring intelligibility involves a comparison of the listener's interpretation of the patient's production with the targeted one. To this end, the speech therapist writes down what he/she perceived and later compares this to the target or - if available - uses multiple choice forms to select the perceived test items [66]. The intelligibility score can than for example be computed

as the number of correctly identified words or phonemes. An advantage of this method is that it can explore different speech pattern deficits. Disadvantages are that this method is more time-consuming and that it sometimes needs multiple persons to evaluate utterances in order to avoid familiarity issues.

4.1.1.4 More profound articulatory investigation

Intelligibility tests are primarily used to determine the severity of the speech disorder. However, speech therapists are also interested in the nature of the disorder to determine the right personal therapy for every patient. Some intelligibility assessments, like the DIA, embody an articulation inventory or phoneme analysis.

4.1.2 Dutch Intelligibility Assessment

The subjective test automated in this work is the DIA test [6], which was specifically designed to measure the intelligibility of Dutch speech at the phoneme level. Each speaker reads 50 consonant-vowel-consonant (CVC) words[1]. The words are selected from three lists: list A is intended for testing the consonants in a word initial position (19 words), list B is intended for testing them in a word final position (15 words) and list C is intended for testing the vowels and diphthongs in a word's central position (16 words). To avoid guessing by the listener, there are 25 variants of each list and each variant contains existing words as well as pronounceable pseudo-words. For each test word, the listener must complete a word frame by filling in the missing phoneme. In case the initial consonant is tested, the word frame could be something like ".it" or ".ol".

The perceptual intelligibility score is calculated as the percentage of correctly identified phonemes. Previous research [6] has demonstrated that the intelligibility scores derived from the DIA are highly reliable. A test with nine speech scientists rating the same 30 speakers was used to investigate the inter-rater agreement. It yielded an intra-class correlation [67] of 0.91 between the ratings of the different raters. To measure the consistency of the ratings, the intra-rater correlation was determined by letting the raters judge all of the 30 recordings twice, with a time interval of more than a month. This lead to an intra-class correlation of 0.93. As the DIA systematically assesses all Dutch phonemes in every word position they may occur in, a qualitative analysis can be performed. Per list, a confusion matrix can be made indicating the differences between targeted and perceived phonemes. These differences can then be accumulated and interpreted in terms of shifts in voicing, manner and/or place of articulation. Previous research [68] showed that inter-rater agreement for phoneme identification however strongly depends on the intelligibility of the speaker. Phoneme errors made by highly intelligible speakers

[1] The empty consonant is also allowed in initial or final position.

were consistently identified by all raters, while the agreement decreased almost linearly with intelligibility.

4.2 Automatic Evaluation

An important issue in intelligibility assessment is that the listener should not be too familiar with the tested speaker since this creates a positive bias. While all other aspects for obtaining a reliable test, like contextual information, can be avoided by choosing the right test material, the listener's knowledge of the speaker or its type of speech can never be totally excluded. In particular, if one wants to use the test for monitoring the efficiency of a therapy, one has to work with different listeners over time. The latter automatically excludes the speaker's therapist as a listener, which is very unfortunate from a practical viewpoint. A possible way to create an unbiased objective listener is to use the computer as a listener. Over the last couple of years there has been a growing interest in trying to apply automatic speech recognition for the automation of traditional perceptual tests and, more generally, to obtain an intelligibility measure using an automatic speech recognizer (ASR). The idea behind using an ASR for intelligibility measurement is quite logical: if the effort a speech therapist has to make to decode a patient's speech is inversely related to his/her intelligibility, then the ease of automatic speech decoding might also point to intelligibility.

Previous research indeed indicated that ASRs can be used for intelligibility measurement. Ferrier et al. [69] experimented with repeated readings of the same passage to a dictation system (Dragon Dictate). In a test on ten dysarthric speakers, they obtained high correlations between mean recognition rate over eight readings and the perceptually measured intelligibility scores. More recently, Vijayalakshmi et al. [70, 71] proved that the phoneme recognition rate is valuable as a measure for intelligibility. Again, only nine dysarthric speakers were tested. A handful of other objective intelligibility assessments have been reported [70, 72–75], but a major limitation of using ASR for intelligibility assessment is that it takes many recordings of pathological speech before the outcome is reliable. As pathological speech corpora are rather sparse, only a few tools are developed that can be used by speech therapists in their daily therapy. Two tools for evaluation of pathological speech will be discussed in more detail in the next subsections.

4.2.1 CFDA

A first tool for the evaluation of pathological speech is the Computerized Frenchay Dysarthria Assessment (CFDA, [4]). This tool is based on the original Frenchay Dysarthria Assessment (FDA, [76]), a perceptual test developed with the aim to fully characterize the speech of a dysarthric speaker. The FDA and CFDA are thus

designed specifically for dysarthric speakers, not for pathological speakers in general. The FDA consists of 28 subtests, each evaluating one aspect of the patient's speech: reflexes (coughing, swallowing, etc.), respiration, and functioning of lips, jaws, tongue, soft palate and larynx are measured. As a last test, intelligibility is also investigated on three levels: word, sentence and conversation.
The CFDA assesses word and sentence intelligibility using the same test material as the FDA. For the intelligibility test at word level, 12 test words are randomly selected from a 50 word database. The first two words are practice items, the other 10 are the actual test words. The words are separately presented on a PC-screen. For the sentence intelligibility 12 sentences are randomly selected from a set of 50.
While most objective intelligibility assessments use word recognition, this test interestingly adopts a forced alignment strategy: a HMM-based ASR (see Section 5.11), trained on speech of non-disabled (normal) English speakers, lines up the patient's utterance with the target words or sentences. The ASR returns a log-likelihood score describing how well the utterance matches the target text, which is called the *goodness of fit* (GOF). Comparison between the GOF of normal speakers and that of dysarthric speakers renders a distance measure that appears to be (cor)related to the perceptually evaluated intelligibility of these patients by means of the FDA test. However, the precise nature of the (cor)relation was not determined. Moreover, only five dysarthric speakers were tested, which does not allow the formulation of any statistically significant result. The author also noticed that the GOF scores depend on the type of utterance: longer and more acoustically complex utterances tend to return lower GOF scores. More research about normalization and standardization of these scores is necessary.
Although the intelligibility aspect of the CFDA needs further elaboration, the tool deserves attention because of its use of forced alignment instead of word recognition, and also because of its versatility and its high level of automation: 12 of the 28 subtests of the FDA have been automated with varying success, amongst which the laryngeal functions and some of the palatal, lip and tongue functions. The other subtests rely more upon visual assessment and/or require information gathered over an extended period of time [4]. Objective assessment helps the speech therapist to make the right diagnosis of a patient. Moreover, as soon as all analysis results are available, the program automatically derives the specific type of dysarthria.

4.2.2 PEAKS

A second tool I wish to review is the Program for Evaluation and Analysis of all Kinds of Speech disorders (PEAKS) [3]. This software package has been developed by the Pattern Recognition Group of the University of Erlangen to manually

and automatically evaluate a variety of voice and speech disorders. It offers a recording and analysis environment for both perceptual and automatic evaluation and it is accessible via internet. To use this tool, the user only needs a PC or laptop with a web browser, a head set, a sound card and an up-to-date Java runtime environment. The user can then choose between a number of possible recording options, amongst which the "Nordwind und Sonne" passage for adults, which is a phonetically balanced text composed of 108 words (71 disjunctive) and containing all phonemes of the German language. The text is frequently used in speech therapy [7] in German speaking countries (see Appendix B for the full text).

Starting from this test, perceptual and automatic evaluation can be performed. The automatic intelligibility score is derived by using an HMM-based ASR which was trained on non-pathological German speech.

To test the performance of this tool, a group of 41 laryngectomees with TE-speech were perceptually evaluated by 5 voice professionals. Per patient, the 5 intelligibility scores, expressed on a 5-point Likert scale [65] were averaged to one reference score. Previous results showed high correlations between the reference and automatic scores [3], which are in the range of the inter-rater agreement of the 5 raters. The results and methodology will be discussed in more detail in Chapter 8.
For young children, a pictogram test can be used, namely the "Psycho-Linguistische Analyse Kindlicher Sprech-Störungen" (Psycho-Linguistic Analysis of Children's Speech Disorders (PLAKSS, [77]). This test consists of 99 pictograms of words to be named by the child. These words cover all German phonemes in different positions. Again, perceptual and automatic evaluation can be performed online. In this case, the automatic intelligibility score is derived by using an HMM-based ASR which was trained on "normal" children's speech as well as adult's speech that was adapted by vocal tract length normalization [78]. The performance of this part of the tool was tested on recordings of 31 children with cleft lip and palate, which were also rated by 5 voice specialists. Similarly as for the TE-speakers, one average rating on a 5-point scale was used as a reference score and the correlations between the reference and the automatic score were almost as good as the inter-rater agreement between the 5 raters. The methodology and results will be discussed further in Chapter 8.
While the main aim of the PEAKS tool is to assess intelligibility of patients with speech disorders, attempts are made to further extend its scope. So far, the only extra feature implemented so far is a visualization tool, positioning the speaker against a group of pathological speakers using a dimension-reduction algorithm called Sammon Mapping [79].

4.2.3 Need for a Flemish tool

Having presented the current tools for intelligibility assessment, two remarks can be made. The first one is that there are only a few tools available. This striking lack of automatic measures is partly due to the fact that there is still insufficient confidence in the abilities of computer models, making the speech therapists rather rely on perceptual tests. A second remark is that none of the few existing Dutch/Flemish perceptual tests [66] have been automated yet. The rather new DIA test (described in Section 4.1.2), developed in 2006, has proven to be highly reliable, and is therefore a good candidate for automation. During the SPACE-project, I developed an automated test, but before describing my approach, there are some fundamental concepts which need to be discussed. In fact, automating an intelligibility test implies the creation of an intelligibility model. To build such a model, one needs a suitable modeling technique and a database (corpus) on which it can be trained and tested. Modeling techniques will be discussed in Chapter 5, the used databases will be described in Chapter 6.

5

Speech recognition basics

In the field of speech recognition, researchers exploit a number of machine learning (ML) techniques to extract valuable symbolic information from a speaker's utterance, but also to characterize the speaker as a speech generator. As the field of ML is very broad, this chapter only presents the ML techniques which were used in this thesis. Subsequently, the principles of automatic speech recognizers (ASR) are briefly explained.

Machine learning techniques always start from a set of examples, called the dataset. Each example is represented by a number of characteristics called *features*. In case of *supervised learning*, a target value or a symbolic interpretation (a label/a class) is available and is also used during the learning of the model. In this case, a model can be built to predict the targeted outcome as well as possible. This outcome can be quantitative, i.e. a range of values, or qualitative, i.e. belonging to one of several categories. In the first case, *regression* techniques are appropriate, while in the latter case *classification* techniques are in order. If no targeted outcome is available, so-called *unsupervised learning* techniques can be applied to expose structure in the data which can enhance insight in the properties of specific examples.

Models built with statistical learners need to be evaluated and compared. Sections 5.2 and 5.3 will deal with evaluation procedures and some performance measures.

5.1 Symbols and notations

In the next sections, fixed values like dimensions will be denoted by a capital letter, like e.g. N. Corresponding indices will be denoted by corresponding small letters, e.g. $n = 1, ..., N; m = 1, ..., M$. Vectors will be be denoted by the capital letters X and Y, while matrices are denoted with bold capital letters, e.g. $\mathbf{X}$. Parameters will be denoted by Greek letters α, β etc. Small bold Greek letters denote vectors, Capital bold Greek letters denote matrices.
Often, we will start from a dataset of N samples. Every sample $n = 1, ..., N$ consists of M features and can thus be represented as a (row) vector X_n with M elements. These vectors can be collected in a NxM matrix $\mathbf{X}$. In case of supervised learning, the row vector Y of length N contains the desired outputs for the examples in $\mathbf{X}$ and $\hat{Y}$ denotes the prediction of Y.

5.2 Evaluation techniques

When searching for an appropriate model for a given task, it is important to investigate whether this model performs well on data that were in no way used during the model development, and which are therefore called independent test data. It is rather easy to design a model which completely fits to the training data. This model however will lack any generalization power. Hastie et al. [80] show that more complex models tend to generalize poorly as more parameters need to be tuned on the same amount of training data. That is why testing on examples that were not involved in the training is needed to expose this phenomenon.

Only a model which performs well on the independent test data is general enough to provide reliable predictions for new examples. It is therefore a standard procedure to split all available data into a training set, a validation set and a test set. If several models are evaluated, those models are built (trained) on the training set and then model selection is achieved on the validation set. Model selection aims at identifying the model which is expected to yield the best performance on an independent test set [80].

In many cases, there are insufficient examples to split them into three disjunct parts. Another very widely used approach for estimating the prediction error is then used, called *(K-fold) cross-validation*. For this method, the dataset is split into K equal-sized parts. $K - 1$ parts are used for model training, and the Kth part is used for testing. This is repeated for every possible combination of $K - 1$ parts out of K, leading to K performance estimates, which are then averaged in order to provide the final cross-validation (CV) performance. Usually, K is set to 5, 10 or the number of samples N. This last case is also known as *leave-one-out cross-validation*. One should be careful in selecting the value of K. It is shown in [80] that the higher the size of one part is, the more accurate the

performance prediction is as the K training sets are similar, leading to similar models. However, the similar models will also lead to a high variance on the K performance predictions. On the other hand, low values for K will suffer from a negative bias on the performance prediction as models are built with less data which might not represent the diversity of the whole dataset, but the results will be less variable as the K test sets are larger. Overall, recommended choices for K are 5 or 10 [80].

5.3 Performance measures

In supervised learning, several performance measures exist. For regression, the Pearson Correlation Coefficient (PCC) and the Root Mean Squared Error (RMSE) between the computed and the targeted outputs are very popular. Starting from the dataset $\mathbf{X}$, the targeted outputs $\mathbf{Y}$ and the predictions $\hat{\mathbf{Y}}$, the PCC is defined as:

$$PCC = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})(\hat{Y}_i - \bar{\hat{Y}})}{\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}\sqrt{\sum_{i=1}^{N}(\hat{Y}_i - \bar{\hat{Y}})^2}} \tag{5.1}$$

The RMSE is defined as the square root of the mean squared prediction error:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2} \tag{5.2}$$

Both measures offer a different view on the results. While PCC quantifies the linearity of the relationship between the outcomes and their predictions, RMSE quantifies the absolute prediction error. RMSE has the advantage of being directly interpretable. In case the discrepancies (errors) are normally distributed, 67% of the computed scores lie closer than the RMSE to the measured (correct) scores. During this research, we observed that if a model is designed to cover a large range of outcomes, and if it is evaluated on a subgroup with a smaller subrange, the PCC can become quite low for this subgroup even though the errors remain acceptable. This happens when the rankings of the samples of this group along the outcomes and their predictions respectively are significantly different. The RMSE results were found to be much more stable across subgroups [13].

In case of classification, the misclassification error rate, defined as the percentage of misclassified samples, and the mean class misclassification rates, defined as the percentage of misclassified samples per class, were used.

5.4 Linear Regression

The simplest prediction model in case of regression is the linear model. It presumes a linear relation between the observations and the outcomes. The parame-

ters $\boldsymbol{\beta} = \beta_0, .., \beta_M$ need to be tuned to predict Y according to

$$\hat{Y} = \mathbf{X}\boldsymbol{\beta}, \tag{5.3}$$

in which β_0 denotes the intercept of the model and the rows of $\mathbf{X}$ also contain an extra constant of 1 to condense the formulae. Tuning of the parameters depends on how the error criterion is set. With the least squares method, the residual sum of squares $RSS(\beta)$ is minimized:

$$RSS(\boldsymbol{\beta}) = ||(Y - \mathbf{X}\boldsymbol{\beta})||^2, \tag{5.4}$$

which, in case $(\mathbf{X}^T\mathbf{X})^{-1}$ exists, eventually leads to the solution

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y, \tag{5.5}$$

or

$$\hat{Y} = \mathbf{X}\boldsymbol{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y = \mathbf{H}Y, \tag{5.6}$$

in which $\mathbf{H}$ is called the hat matrix (as it puts the "hat" on Y). This matrix projects Y orthogonally onto the space spanned by the observation vectors in $\mathbf{X}$. Its trace defines the model complexity, being the number of parameters to be tuned [80]. When the features (columns of $\mathbf{X}$) are not independent of each other (and thus creating redundancy), the matrix $\mathbf{X}$ is not full-rank. As a consequence, $(\mathbf{X}^T\mathbf{X})$ will be singular and the coefficients $\boldsymbol{\beta}$ are not uniquely defined anymore. A simple solution to this problem is dimensionality reduction.

Dimensionality reduction. The problem of redundancy in the feature set can be solved in- or outside the regression program. Apart from the redundancy problem, the size of the feature set can be a source of overfitting as it directly determines the model complexity. Dimensionality reduction involves the derivation of a subset of uncorrelated features from the original feature set.

This derivation can be done by means of *principal component analysis* (PCA), which projects the original (correlated) features onto a smaller set of uncorrelated features, called the principal components [81]. Transforming the feature space however has the drawback that the new PCA-dimensions do not always have the same simple physical interpretation as the original dimensions might have. Therefore, one could consider feature selection instead of feature transformation. Some methods, like ANOVA [80], search for the features which best explain the outcome variable. These methods however do not always retain the best feature set as interactions between features are not taken into account or only to a limited extent. Therefore, another approach is to simply build a model for every possible feature subset, to test these models on the validation set and to select the best one. When the feature set is large, it is computationally inefficient to perform such an exhaustive search and other search methods need to be applied. The simplest

search algorithm is the greedy forward or backward selection. The forward procedure starts with the best combination of a small number of features and adds one feature (the best) at the time. The backward procedure starts with all the features and removes one feature at the time. A variety of smarter feature selection techniques exist, like e.g. the use of genetic algorithms [82], but again these techniques are way more computationally expensive.

5.5 Regression and classification trees

Another simple though nonlinear modeling technique is a decision tree. In its simplest form, a tree partitions the feature space into hyperrectangles in which the fitted outcome has 1 particular value. More complex forms of decision trees are discussed in [83]. Decision trees can be used for both regression and classification.

Growing trees The hyperrectangles in the feature space are obtained by consecutively splitting example sets according to the value of one feature. The following process is repeated recursively:

- Per feature $m = 1, ..., M$, do:
  - For all t between the minimum and maximum value of $\mathbf{X}_{im}$, $i = 1, ..., N$, do:
    - Split the set of examples into two groups: observations i with $\mathbf{X}_{im} \leq t$ and observations i with $\mathbf{X}_{im} > t$.
    - All observations of one region obtain the same prediction, being the mean outcome (regression) or most likely class label (classification) of all observations in that region.
    - Determine the predictive power of this model in terms of RMSE or PCC in case of regression and the misclassification error in case of classification.
  - Establish the *split point* t_m that leads to the highest prediction power over all possible values of t.
- Find over all $t_m, m = 1, ..., M$ the optimal split for this iteration.

This process is repeated in a recursive way. The recursion is completed when splitting no longer improves predictive power. The final regions in the tree are called the *leafs*.

Pruning trees With each split, the complexity of the decision tree augments. While growing a tree the risk of overfitting therefore rapidly increases. Therefore it is important to prune the tree. One approach could be to only add split points if this enhances the predictive power to the validation set. However, a seemingly worthless split might lead to a very good split further in the process. Therefore, the preferred strategy is to first grow the whole tree and to prune afterwards [80].

5.6 Combining weak classifiers

Both linear regression and decision trees are simple but powerful methods. They are however known to be very sensitive to noise in the data and to produce highly variable results [80]. Moreover they can easily overfit the training data, which makes them in a sense weak learners. Bagging and boosting are two powerful methods to overcome these shortcomings. They both aggregate many weak predictors to obtain one stronger model with lower variance.

Bagging The idea behind bagging is that averaging out the results of many weak predictors leads to a more robust model. Therefore, many weak models of the same kind, such as decision trees, are created and later aggregated. In order to do so, B new training sets, each with the same size N, are created by randomly drawing samples from the original set. Samples are drawn with replacement, which implies that one example can occur more than once in one training set. The idea behind this method is that we randomly draw samples of size N from the empirical distribution of the data, of which the original dataset is only a part. This empirical distribution is discrete and puts an equal weight on each observation in the dataset. By sampling with replacement, it is likely that some examples will be repeated in each new dataset. If N is large, every set is expected to have $0.632\text{x}N$ unique examples and $0.268\text{x}N$ duplicates [80].

A weak model is trained on every training set. The final prediction for a sample is then the average of all predictions (in case of regression), or the majority vote (in case of classification).

Boosting Boosting originates from the same idea as bagging, but the way of combining the weak models to stronger ones makes this method fundamentally different. Boosting algorithms start with applying one weak predictor $G_1 : \mathbf{y}_i = G_1(\mathbf{x}_i), i = 1, ..., N$. The observations are then weighted according to the quality of their prediction: the better the quality, the lower the weight of the observation. With this new dataset, in which the observations are weighted, the same learner is used to create another model. Again, after training the model, the observations are weighted according to the results. This iterative process is repeated K times, creating a sequence of models $G_1, ..., G_K$. The K models are then combined in

a weighted manner (weighted majority vote or weighted average) to produce the final model $G(\mathbf{x}_i)$ [80].

Random Forests Random forests (RF) is an improved implementation of the bagging principles for decision trees. Like in bagging, RF grows a large number of trees (hence a forest). The difference with bagging lies in the provided feature set: each node in the tree can only split according to a randomly selected subset of the features. It can be shown that this random feature selection prevents overfitting to the training data [80], which is a major advantage as it gives rise to stable models with reliable predictions.

Ensemble Linear Regression incorporating feature selection One regression technique which can be used on a small data set is ensemble linear regression (ELR), which combines the low model complexity of linear regression with a bagging strategy [84]. The latter boosts the predictive power by using many simple models (linear regression models in this case) which are each trained on a different random subset of the training data. For the training of our ELR model, ten random divisions of the training set into two equally large parts are created: one part for estimating the regression coefficients and the other for assessing the model. If the experiments involve a large number of features and as every division will only comprise a very restricted number of speakers, some feature selection procedure is indispensable. Every single model is created by adopting a greedy forward feature selection procedure which starts with the feature leading to the best performance on the validation part of the random split and continues to add features as long as that performance rises.

The ten models emerging from the ten training set divisions are then combined into one single model by just averaging the regression coefficients of these models. This final model is then evaluated on the test set.

5.7 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) and the related Fisher's linear discriminant analysis are classification techniques which aim at finding the set of hyperplanes in the feature space which maximize the separation of K classes. While LDA assumes Gaussian class distributions with equal covariance matrices for all classes, Fisher's approach does not use this constraint to find the linear transformation $\mathbf{Z} = \mathbf{a}^T\mathbf{X}$ of the feature space according to which the variance of the class centers (between-class variance) is maximized relative to the within-class variance. This latter is the weighted average over all classes of the variance of the observations around their class means. The weights are equal to the class probabilities. Starting

from a dataset $\mathbf{X}$, the between-class variance $\mathbf{\Sigma}_B$ and the within-class variance $\mathbf{\Sigma}_W$ can be written as follows:

$$\mathbf{\Sigma}_B = \sum_{i=1}^{K} \pi_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T, \tag{5.7}$$

with π_i the prior probability of class i, $\boldsymbol{\mu}_i$ the mean value of $\mathbf{X}$ in class i and $\boldsymbol{\mu}$ the overall mean value of $\mathbf{X}$, and

$$\mathbf{\Sigma}_W = \sum_{i=1}^{K} \pi_i \mathbf{\Sigma}_i, \tag{5.8}$$

where $\mathbf{\Sigma}_i$ denotes the covariance matrix of class i. The matrix $\mathbf{a}$ of Fisher's problem can then be found by maximizing the function $A(\mathbf{a})$:

$$A(\mathbf{a}) = \frac{\mathbf{a}^T \mathbf{\Sigma}_B \mathbf{a}}{\mathbf{a}^T \mathbf{\Sigma}_W \mathbf{a}}, \tag{5.9}$$

which can be solved as an eigenvalue problem, with a the eigenvector corresponding to the largest eigenvalue of $\mathbf{\Sigma}_W^{-1}\mathbf{\Sigma}_B$ [80]. All eigenvectors together determine a new feature space of maximally $K-1$ dimensions. This makes LDA very interesting in case the feature space of $\mathbf{X}$ is much larger than K as it reduces the feature space, which is useful for visualization purposes.

5.8 Multi Layer Perceptrons

A totally different kind of statistical learner is the multi layer perceptron (MLP). This special type of neural network can be seen as a nonlinear regression or classification model, in which the outputs of each layer are computed as a function of a linear combination of the outputs of the preceding layer. A typical architecture of an MLP is depicted in Figure 5.1. This MLP consists of three layers, being the input layer, the hidden layer and the output layer. The input layer consists of the M input nodes, one for each component of the M-dimensional input vector X_n. Each input node $m, m = 1, ..., M$ can be connected to each hidden node $q, q = 1, ..., Q$ by means of an arc carrying an adjustable weight α_{mq}. Similarly, each hidden node q can be connected to each output node $p, p = 1, ..., P$ by means of an arc carrying a weight β_{qp}, leading to the estimation $\hat{Y}_n$ of the targeted p-dimensional output Y_n. Every node q of the hidden layer computes the value $F_q(X_n)$ and every node p of the output layer computes $\hat{Y}_{np}$ according to:

$$F_q(X_n) = \frac{1}{1 + e^{-a_{nq}}} \qquad \text{with} \qquad a_{nq} = \sum_{m=1}^{M} \alpha_{mq} X_{nm}, \tag{5.10}$$

$$\hat{Y}_{np} = \frac{1}{1 + e^{-b_{np}}} \qquad \text{with} \qquad b_{np} = \sum_{q=1}^{Q} \beta_{qp} F_q(X_n). \tag{5.11}$$

The MLP can be designed to predict the classification or regression problem as

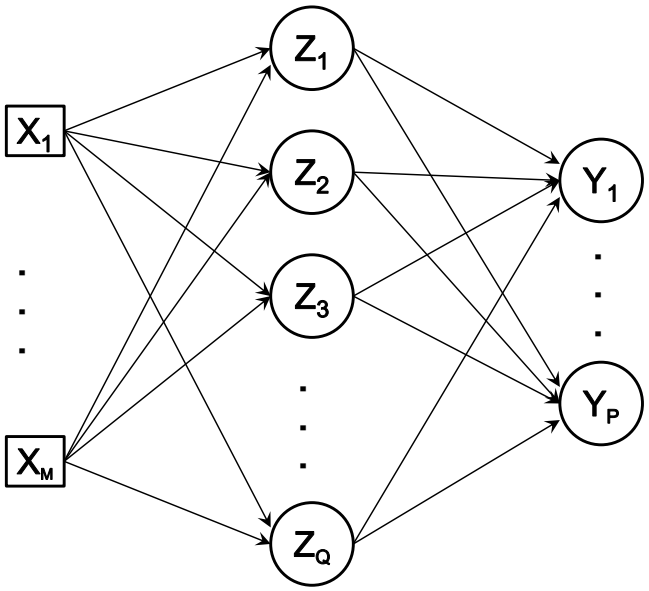

Figure 5.1: Schematic of a multi layer perceptron with one hidden layer. Every circle denotes a computing node, resulting in the value marked within the node.

well as possible. First of all, the number of hidden nodes Q can be chosen. This number will determine the model complexity as it confines the number of possible interconnections (arcs) in the network. Secondly, the interconnection scheme can be defined to possibly restrict the number of arcs. These two parameters can be chosen by the user and they define the number of weights to be tuned.

The weights are trained to minimize the sum of square errors $R(\alpha, \beta)$ between the computed and the targeted outputs:

$$R(\alpha, \beta) = \sum_{n=1}^{N} \sum_{p=1}^{P} (Y_{pn} - \hat{Y}_{pn})^2, \tag{5.12}$$

with N the number of training instances. The minimization of this function of α and β is usually done by applying the error back-propagation (EBP) algorithm [80].

5.9 Support Vector Machines

A Support Vector Machine (SVM) is a binary classifier that uses the optimal separating hyperplane between two classes $Y = 1$ and $Y = -1$. Such a hyperplane can be represented by an equation like

$$f(X) = X\boldsymbol{\beta} = 0. \tag{5.13}$$

In case $f(X) > 0$, X is assigned to class $Y = 1$, otherwise X is assigned to class $Y = -1$.

In case the data are separable, $\boldsymbol{\beta}$ is scaled so that the hyperplanes H_1 for which $X\boldsymbol{\beta} = 1$ and H_{-1} for which $X\boldsymbol{\beta} = -1$ form the class borders. Often there will be many possible hyperplanes fully separating the classes. SVMs therefore define the optimal separating hyperplane as the one maximizing the margin between H_1 and H_{-1}, as is depicted in Figure 5.2a.

As the distance between H_1 and H_{-1} is equal to $\frac{2}{||\boldsymbol{\beta}||^2}$, the optimal hyperplane can be found by solving

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \, ||\boldsymbol{\beta}||^2 \tag{5.14}$$

subject to

$$Y_n \frac{X_n\boldsymbol{\beta}}{||\boldsymbol{\beta}||} \geq 1; \qquad n = 1, ..., N. \tag{5.15}$$

With these constraints, the solution is uniquely defined [85]. It can be shown [80] that the direction of the optimal separating hyperplane can be found as a linear combination of those datapoints for which Equation (5.15) is exactly met, meaning the points which are on the boundary of the margin. Those points are called the *support vectors*.

In most cases, the two classes are not fully separable, like in Figure 5.2b. To this end, *slack variables* $\xi = (\xi_1, ..., \xi_N)$ are introduced to create a so-called *soft margin* [80]. The variable ξ_n allows datapoint X_n to fall at the wrong side of its margin. Fixing the sum of the slack variables, $\sum_{n=1}^{N} \xi_n = T$, defines the maximal allowed number of misclassified training points T. The optimal hyperplane is then found by optimizing the following soft margin problem:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \; [\frac{1}{2} \, ||\boldsymbol{\beta}||^2 + C \sum_{n=1}^{N} \xi_n] \tag{5.16}$$

subject to

$$Y_n(X_n\boldsymbol{\beta}) \geq 1 - \xi_n; \qquad n = 1, ..., N, \tag{5.17}$$

$$\xi_n \geq 0; \qquad n = 1, ..., N. \tag{5.18}$$

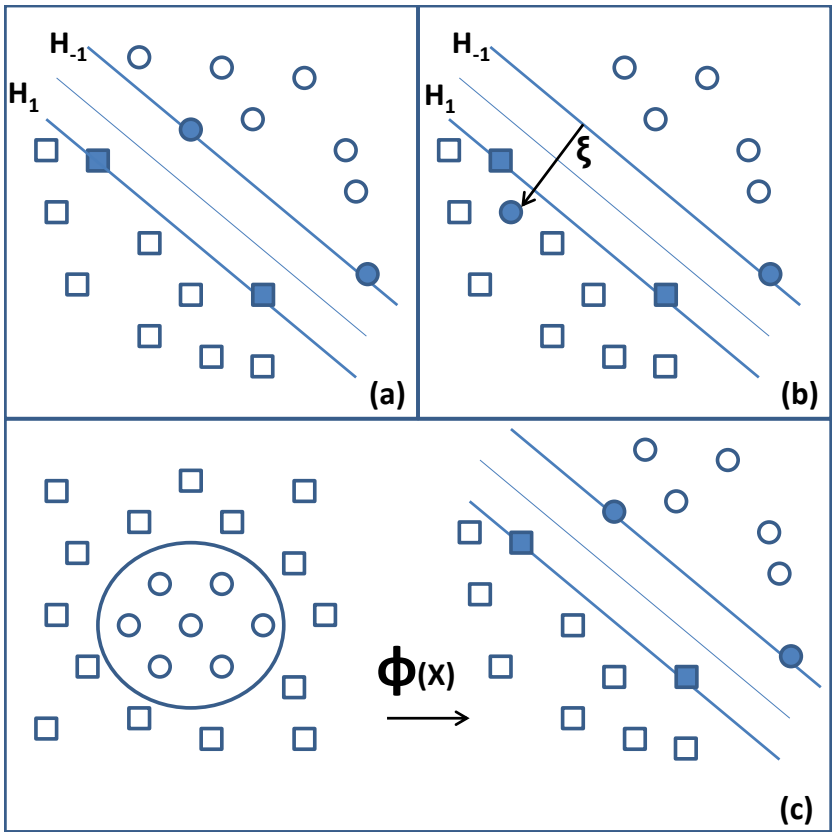

Figure 5.2: Schematic of a SVM. Figure a depicts the case of linearly separable classes: squares and bullets can be separated linearly. Filled bullets/squares denote support vectors. Figure b depicts the case of non-separable classes, introducing the slack variable ξ. Figure c denotes the case of nonlinearly separable classes. Using a Gaussian kernel transformation $\Phi(X)$, the classes are transformed into the linearly separable case.

In these equations, the parameter C, determining the penalty of misclassification, can be chosen by the user. It can be shown [80, 85] that the solution for $\boldsymbol{\beta}$ can be found by maximizing the function

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmax}}(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j Y_i Y_j X_i^T X_j \tag{5.19}$$

subject to

$$\sum_{i=1}^{N} \alpha_i Y_i = 0 \tag{5.20}$$

$$0 \leq \alpha_i \leq C, \tag{5.21}$$

leading to

$$\hat{\boldsymbol{\beta}} = \sum_{i=1}^{N} \hat{\alpha}_i Y_i X_i. \tag{5.22}$$

As shown in [80], only those α_i for which constraint (5.17) is exactly met are not equal to zero. The corresponding X_i are called the support vectors. The resulting $\hat{\boldsymbol{\beta}}$ is thus a linear combination of the support vectors of the non-separable class problem, all lying on the margin ($\hat{\xi}_i = 0$) or on the wrong side of the margin ($\hat{\xi}_i > 0$).
It can be noticed from Equation (5.19) that in the formulation of the problem the training data only appear as dot products. This opens the possibility to generalize the SVM classifier to non-linear separation boundaries. To this end, the training data are first subjected to a non-linear transformation $\Phi(\mathbf{x})$. In this new and possibly infinite dimensional feature space, the optimal hyperplane is searched as explained before. This is depicted in Figure 5.2c. As the training data only appear as dot products in Equation (5.19), the transformation $\Phi(\mathbf{x})$ does not have to be known exactly, as only the *kernel* function $K(\mathbf{x}_i; \mathbf{x}_j) = \Phi(\mathbf{x}_i).\Phi(\mathbf{x}_j)$ is needed to solve the equations, reducing the possibly infinite dimensional transformation to a finite-dimensional criterion [80, 85]. Some frequently used kernels are:

- The linear kernel : $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \mathbf{x}_j$
- The polynomial kernel : $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \mathbf{x}_j + 1)^p$
- The Gaussian radial basis function kernel : $K(\mathbf{x}_i, \mathbf{x}_j) = e^{(} - \frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma^2})$

As the solution is only dependent on the support vectors, the complexity of the resulting SVM is determined by the number of support vectors rather than the dimensionality of the transformed space.

Some parameters are user-defined and can be determined by an external cross-validation loop: e.g. the number of allowed misclassifications C and, in case of a radial basis function (RBF) kernel, $\gamma = \frac{1}{2\sigma^2}$, are important parameters. Usually, these are determined by a grid search or by using an alternative method like [86].

Support Vector Regression The ideas behind SVM can be adapted for regression. While SVM classifiers determine the boundary with the largest margin separating two classes, Support Vector Regression (SVR) determines the regression hyperplane around which most observations lie closer than a distance ϵ. Like datapoints on the correct side of the decision boundary do not play a role in the optimization criteria for SVMs, an ϵ-insensitive error measure $V_\epsilon(r)$ is defined which is equal to zero in case the distance $|r| = Y_i - \hat{Y}_i$ between a datapoint and the optimal hyperplane is smaller than ϵ:

$$V_\epsilon(r) = max(0, |r| - \epsilon). \tag{5.23}$$

Starting with the linear case, the linear regression model $f(X) = X\boldsymbol{\beta}$ is then solved by SVR by optimizing the problem

$$\hat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta}}[\sum_{i=1}^{N} V_\epsilon(Y_i - f(X_i)) + \frac{\boldsymbol{\lambda}}{2}||\boldsymbol{\beta}||^2]. \tag{5.24}$$

It can be shown [80] that the minimization of this equation leads to a function $\hat{f}(X)$ of the form

$$\hat{f}(X) = \sum_{i=1}^{N} \hat{\alpha}_i X^T X, \tag{5.25}$$

with $\hat{\alpha}_i$ only different from zero for a subset of the datapoints X_i, called the support vectors. It can be noticed that the solution only depends on the datapoints through a dot product, which again makes it possible to expand the use of SVR to nonlinear regression using kernels like described above. A schematic of Support Vector Regression can be found in Figure 5.3.

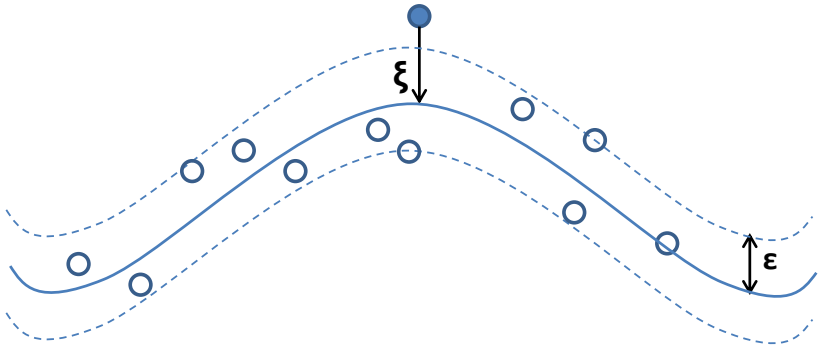

Figure 5.3: Schematic of a nonlinear SVR. Using a RBF kernel, this problem will be reduced to a linear SVR. The middle line denotes the regression line. ξ denotes the slack variable, defined like in case of an SVM.

5.10 Gaussian Mixture Models

Gaussian Mixture Models (GMM) are often used in speech processing to model the acoustic properties of a class of data, e.g. one phone or one (type of) speaker. Samples X of such a class $C_j, j = 1, ..., K$ with K equal to the number of possible classes, are then characterized by a probability density function $p_\theta(X, C_j)$, consisting of a weighted sum of M Gaussian densities:

$$p_\theta(\mathbf{x}|C_j) = \sum_{i=1}^{M} \mathbf{W}_{ij}\mathcal{N}_{ij}(X) = \sum_{i=1}^{M} \mathbf{W}_{ij} \frac{1}{(2\pi|\boldsymbol{\Sigma}_{ij}|)^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{M}_{ij})^T \boldsymbol{\Sigma}_{ij}^{-1}(\mathbf{x}-\mathbf{M}_{ij})}. \tag{5.26}$$

Every Gaussian $\mathcal{N}_{ij}(X)$ of class C_j is characterized by its weight $\mathbf{W}_{ij}$, mean $\mathbf{M}_{ij}$ and covariance matrix $\mathbf{\Sigma}_{ij}$. These parameters $\theta = \{\mathbf{W}, \mathbf{M}, \mathbf{\Sigma}\}$ can be optimized during training of the GMM by applying Maximum-Likelihood Estimation (MLE) [80]. This method aims at maximizing the log likelihood function $ll_j(\theta, X)$, which describes the joint probability of all observations belonging to the modeled class C_j with cardinality N_j under the trained model:

$$ll_j(\theta, X) = \sum_{i=1}^{N_j} log(p_\theta(X|C_j)). \tag{5.27}$$

The maximization of the log likelihood function can be achieved by using the iterative *Expectation-Maximization* algorithm [80]. Starting from first estimates, θ is iteratively adapted to increase $ll_j(\theta, X)$ until convergence. The resulting GMMs can form the base of HMMs, which will be discussed in the next Section.

5.11 Hidden Markov Models

A Hidden Markov Model (HMM) is a finite state machine that can generate a sequence of observations according to a stochastic Markov process. It is typically represented by a graph in which each node denotes a state and each arc denotes a transition, as depicted in Figure 5.4. Self-loops are also possible transitions.

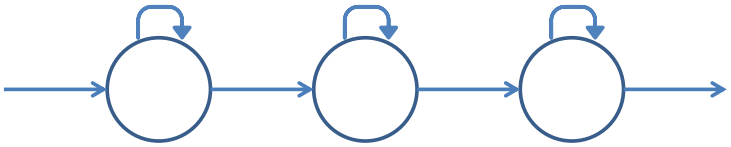

Figure 5.4: Typical structure of an HMM. Each node denotes a state and each arc denotes a transition.

Beginning in an initial state, the observation sequence is generated by moving from one state to another, following a path along the sequence of states $\mathbf{S} = s_1, ..., s_N$. Typically for this process is that the probability of being in a state at time n only depends on the previous state: $P(s_n|s_1, ..., s_{n-1}) = P(s_n|s_{n-1})$. The process of moving through these states is characterized by the transition probabilities $a_{ij} = P(s_n = j|s_{n-1} = i)$ between states i and j. The part of the process that makes the Markov process a *hidden* Markov model, is the fact that the states $s_1, ..., s_N$ are not observable. In state $s_n = j$, there is a certain emission probability $b_j(X)$ of generating the (visible) observation X. Based on this formulation, the probability $P(\mathbf{X}|\lambda, \mathbf{S})$ according to a model with parameters $\lambda = a_{ij}, b_j, \pi$ of

a given observation sequence $\mathbf{X} = X_1, ..., X_N$ of length N to be generated by a sequence of states $\mathbf{S}$ can be computed as [87]:

$$P(\mathbf{X}, \mathbf{S}|\lambda) = P(\mathbf{S}|\lambda).P(\mathbf{X}|\mathbf{S}, \lambda) \tag{5.28}$$

If we now assume that observations are independent from each other, and since the probability of being in a state only depends on the previous state, we can write

$$P(\mathbf{S}|\lambda) = \pi_{s_0} \prod_{i=1}^{N} a_{s_{i-1},s_i} \tag{5.29}$$

$$P(\mathbf{X}|\mathbf{S}, \lambda) = \prod_{i=1}^{N} b_{s_i}(\mathbf{x}_i) \tag{5.30}$$

where π_{s_0} denotes the prior probability of starting in state s_0. The total probability of the observation sequence $\mathbf{X}$ according to the model is obtained by summing $P(\mathbf{X}, \mathbf{S}|\lambda)$ over all possible state sequences $\mathbf{S}$, resulting in

$$P(\mathbf{X}, \mathbf{S}|\lambda) = \sum_{\mathbf{S}} \pi_{s_0} \prod_{i=1}^{N} a_{s_{i-1},s_i}\, b_{s_i}(\mathbf{x}_i) \ . \tag{5.31}$$

In speech recognition, HMMs are typically used to model phonemes or words. Obviously, it is important to know the most probable state sequence as this will provide us with the most likely phoneme or word segmentation. This state sequence is usually found using the *Viterbi* algorithm [87].

Training of an HMM. Before an HMM can be used, its topology needs to be defined and its model parameters λ have to be trained. Typically, the HMM of a phoneme consists of 3 states, which can be traversed from left to right without the possibility of skipping a state, like in Figure 5.4. Self-loops are necessary. HMMs for words consist of the concatenation of the HMMs for the phonemes constituting the word.

Starting from a large speech corpus of so-called training examples of which the phonemic transcriptions are known, the model parameters can be estimated during training.

5.12 Domain Adaptation

During my research, I was often confronted with rather small datasets. Creating models when data are scarce implies that the resulting models have to be of a low complexity. Fortunately, sometimes the small dataset is part of, or is related to, a larger dataset, on which a more robust and complex model can be trained, which is

usually called a Universal Background Model (UBM). In that case, domain adaptation (DA) can be used to create a model for the scarce dataset (in-domain data) by adapting the model for the larger dataset (out-of-domain data).

In case of GMMs, the adaptation is often done by maximum a posteriori estimation of probability density functions, called MAP adaptation [88]. The advantage of using domain adaptation here is not only to create robust models based on much training data, but also to describe every model in terms of its deviations from the UBM.

Apart from DA techniques such as MAP adaptation, some very simple alternatives are established to tackle the DA problem. In [89], four simple methods are described to bias or adapt a model towards the in-domain data:

- Train a model on out-of-domain as well as in-domain data, but increase the weight of the in-domain data. This way, the model will be biased towards the in-domain data.
- Train an in-domain and an out-of-domain model and take the weighted sum of both as the final model.
- Use the predictions of an out-of-domain model as an extra feature in the feature space in which to create an in-domain model. This can be seen as a cascaded approach.
- Adaptation can also be performed by feature augmentation. Instead of starting from the original feature space with dimension M, a new feature space is created with dimension $3M$. The transformation rule is very simple: if a sample $\mathbf{x}_{old} = (x_1, ..., x_M)$ is part of the in-domain data, the transformed sample $\Phi(\mathbf{x})$ will be defined as the 3M-dimensional vector $\Phi(\mathbf{x}) = (\mathbf{x}, \mathbf{0}, \mathbf{x})$, consisting of a concatenation of the original features, followed by M zeros, followed by again the original features. If the sample is an out-of-domain sample, the transformation will insert M zeros behind the concatenation of twice the original feature vector: $\Phi(\mathbf{x}) = (\mathbf{x}, \mathbf{x}, \mathbf{0})$. The key idea between this is that the in-domain and out-of-domain models share some features and do not share some others. Building a model in this feature space can then be seen as a mixture of a general model, an out-of-domain model and a in-domain model with weights chosen by the statistical learner.

5.13 Automatic Speech Recognition

Although many types of ASRs exist nowadays, they are all based on the same underlying fundaments. Basically, an ASR consists of five modules [90], represented in Figure 5.5: an acoustical feature extractor, a pattern recognizer, a set of acoustic models, a lexicon and a language model.

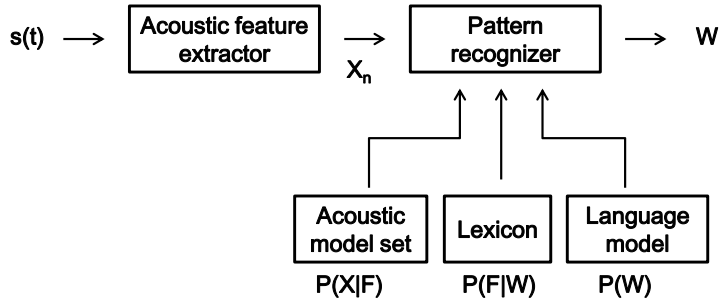

Figure 5.5: Schematic of an Automatic Speech Recognizer. Based on [90]

Acoustical feature extractor. This module transforms the speech signal $s(t)$ into a sequence of acoustic parameter vectors $\mathbf{X}_n$. These parameter vectors are generated by consecutively analyzing small overlapping windows (frames) of $s(t)$. The most popular acoustic features are the Mel-frequency cepstral coefficients (MFCCs) [91].

Pattern Recognizer. This module serves as a search engine to find the most likely word sequence $\hat{\mathbf{W}}$ given the acoustic input vectors $\mathbf{X}$. It aims at maximizing the posterior probability of a word sequence $\mathbf{W}$ given $\mathbf{X}$:

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmax}}\, P(\mathbf{W}|\mathbf{X}) \tag{5.32}$$

According to Bayes' law, this equation can be reformulated as follows:

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmax}}\, \frac{P(\mathbf{X}|\mathbf{W})P(\mathbf{W})}{P(\mathbf{X})} = \underset{\mathbf{W}}{\operatorname{argmax}}\, P(\mathbf{X}|\mathbf{W})P(\mathbf{W}), \tag{5.33}$$

since $P(\mathbf{X})$ does not depend on $\mathbf{W}$. Denoting an arbitrary phone sequence as $\mathbf{F}$, this equation can be specified as follows [11]:

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmax}} \sum_{\mathbf{F}} P(\mathbf{X}|\mathbf{F})P(\mathbf{F}|\mathbf{W})P(\mathbf{W}). \tag{5.34}$$

This equation consists of three factors. $P(\mathbf{X}|\mathbf{F})$, the probability of an acoustic parameter vector sequence given the phone sequence, is modeled by the acoustic models. The second factor, $P(\mathbf{F}|\mathbf{W})$, the probability of a phone sequence given the word sequence $\mathbf{W}$, is modeled in the lexicon. Finally, the probability of a word sequence $P(\mathbf{W})$ is modeled by the language model.

Acoustic Models. Each statistical model estimates the probability that a sequence of acoustic vectors $\mathbf{X}$ can be observed given that a phone $\mathbf{F}$ is realized in a particular phonetic context. The simplest models do not take the phonetic context into account. In that case there is only one model per phone. In modern recognizers however, the models are context-dependent (CD), meaning that there are many models per phone. The number of models can be kept within limits by not creating a model for all possible contexts, but to cluster on the basis of phonological questions [92]. Every leaf of the final (pruned) tree is represented by another acoustical model.

Lexicon. The lexicon can be seen as a pronunciation dictionary. It comprises all the possible words for the given recognition task, together with their phonemic transcription. If a word can be pronounced in several ways, several phonetic transcriptions of that word are incorporated.

Language model. This module defines the a priori probability $P(\mathbf{W})$ of a word sequence $\mathbf{W} = w_1, w_2, .., w_i$ in a specific task. Usually, a so-called *N-gram* language model is used, which models the probability that words w_i occur in succession with the words $w_{i-N+1}, ..., w_{i-1}$. These probabilities are typically derived from large text corpora and smoothed and adapted towards the specific task. The benefit of using a language model is that it reduces the search space of the pattern recognizer as at every point in time only a small percentage of all theoretically possible words are meaningful. This reduction of the search space implies a large speed improvement.

When the language model is restricted to one possible word sequence, the ASR is no longer used as a decoder which tries to reveal the exact words of an unknown message, but rather to find where the different phone(me)s of a known message (with a given transcription) are realized. This process is called *forced alignment*.

6
Flemish databases

During my research, I worked with several databases: one containing Flemish "normal" speech, one with Flemish pathological speech and one with Flemish and Dutch "normal" speech.

6.1 Corpus Gesproken Nederlands (CoGeN)

CoGeN, a corpus of spoken Flemish, was developed by the universities of Leuven and Ghent [93]. It consists of a variety of speech types, ranging from spelled and read isolated words over read continuous speech to simulated man-machine interactions. Recordings took place in an office environment as well as over a telephone. For my research, I only used the office recordings.

Read and spelled words. For this part, 174 Flemish speakers (73 women, 101 men) each spelled a list of 10 words (there were 4 variants of this list), read the 10 digits and read a list of 100 words (again there were 4 variants of this list), leading to 2.16 hours of spelled words and 5.83 hours of read words. The data were provided with broad phonetic transcriptions (meaning no allophonic variation was taken into account). Word boundaries were created using an ASR and the segmentations and phonetic transcriptions were verified and adapted where necessary for all words.

Starting from these verified transcriptions, an ASR was used to create a segmentation at the phone level of the read words, leading to phone labels and time

information for every segment as described in [93]. For 40 speakers, the automatic segmentation was checked manually and corrected where necessary, leading to the so-called manual labels.

Read paragraphs. Each of the 174 speakers read five paragraphs, leading to 7.02 hours of continuous speech. Every utterance was provided with an orthographic transcription of what was actually read. This transcription was then converted into a phonetic transcription using a standard grapheme to phoneme converter [94]. This automatically generated phonetic transcription was again used to create a segmentation at the phone level with an ASR, leading to phone labels with time information for all read paragraphs. For 30 paragraphs, this segmentation and transcription were verified and adapted where necessary. A quantitative study about the followed segmentation procedure and accuracy can be found in [93].

6.2 Dutch Corpus of Pathological Speech (COPAS)

The Dutch Corpus of Pathological Speech [95] was constructed within the framework of the SPACE-project [5]. It served as a resource for the development of the objective intelligibility assessment tool presented in this thesis. COPAS mainly consists of recordings of the DIA test performed by both normal and pathological speakers, but it also contains a variety of other samples like readings of text passages, isolated sentences and spontaneous speech. Since 2010, the corpus is made publicly available through the Dutch Language Union[1].

6.2.1 Speakers and tests

The speakers recorded in COPAS belong to 8 distinct pathological categories, which are shown together with their abbreviation and cardinality (number of speakers of that category) in Table 6.1. For pathologies dysarthria, laryngectomy, hearing impairment and cleft, a wide range of subpathologies are represented. The precise etiology, gender, age etc. of every speaker can be found in the information part of the corpus, together with the exact date and conditions (see section 6.2.3) of his/her recording(s).

As mentioned in the beginning of this section, several tests were recorded, but not every speaker participated in every test. For a majority of the speakers (305), only recordings of the isolated word test (DIA) are available, but for 122 speakers there are also recordings of the standard Dutch text passage “Papa en Marloes” [96] (hereafter denoted with ‘TM’) consisting of 8 phonetically rich sentences (see Appendix B). Other tests include the sustained vowel, diadochokinetic rate, formant transition, readings of 2 sentences and (semi-) spontaneous speech. In my work, I

[1] http://www.inl.nl/en/producten

Speakers	cardinality
Normal (N)	122
Dysarthria (D)	75
Hearing impairment (H)	29
Laryngectomy (L)	30
Cleft (C)	38
Articulation disorders (A)	17
Voice disorder (V)	7
Glossectomy (G)	1
Total	319

Table 6.1: Speakers in COPAS. From [95]

focused on the recordings of the DIA and the passage "Papa en Marloes".

For my experiments, I defined some subsets of COPAS. Concerning the DIA test, I defined a development set of 231 samples, hereafter called 231_DIA. As most of the time only one therapist judged the recordings, we chose for consistency and used only the scores of the speech therapist that judged all samples. As the main purpose is to examine pathological speech, the proposed set comprises the 181 pathological speakers that have performed the DIA. To restrict the influence of the control (normal) speakers, the control group was restricted to 50 samples. If a speaker was recorded multiple times, only the recording of the first session was considered. However, there was one exception to this rule: if the second session also included the reading of a paragraph, that second session was preferred over the first session. Figure 6.1 shows a histogram of the intelligibility scores of these 231 speakers. The 50 control speakers have intelligibility scores between 82 to 100 with an average of 94.3. The pathological speakers' intelligibility range from 28 to 100 with a mean of 81. The histogram shows that there are few examples of very low intelligibility scores. This is probably due to the fact that persons with a very low intelligibility score often suffer from other physical limitations as well, making it difficult to perform the test.

To conclude, a subset of 121 speakers read both TM and the DIA-test. These recordings yielded two datasets: 121_TM and 121_DIA.

6.2.2 Annotations

The recordings were annotated by a skilled speech therapist using the open source program PRAAT [97]. Every annotation file (TextGrid) was organized in tiers: the first tier contains the target text that was presented to the speaker, the second tier represents the orthographic transcription of what the speaker actually read ac-

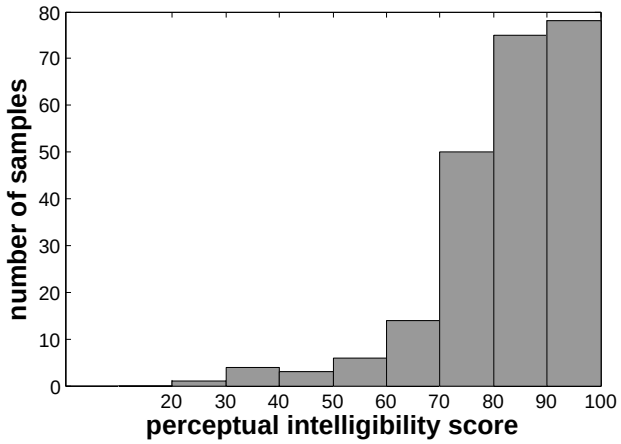

Figure 6.1: Histogram of intelligibility scores in the development set.

cording to the annotator. For the read passages, the tiers were segmented on a sentence level. For the DIA test, the segmentation was on a word level, including non-speech indications for time-intervals not containing any speech of the patient.

6.2.3 Microphone issues

An important issue in COPAS concerns the used microphone. All recordings were made in a quiet clinical setting without sound treated box. Two totally different microphone settings were used: a Sony ECM-717 lying on the table at a distance of about 30 cm from the mouth, and a Shure headset WH20-QTR. The reason for using two different types of microphones is historical. Both microphone types are intrinsically different as they not only show another frequency response but also pick up different parts of the signals due to their different distance to the speaker. Even if both microphones are unidirectional, the Sony microphone lying on the table will capture way more background noise than the head set, simply because its position further from the mouth. This also implies that reverberation could be present in the Sony-recordings. On the other hand, the headset can have the drawback that the signal of plosives formed too close to the microphone can be clipped and that more mouth noises will be audible. For most recordings (249 persons), the recording circumstances are known. For the remaining recordings, the used microphone could not be backtracked by the speech therapist. However, as the two recording techniques differ substantially, a classifier could be trained to retrieve the used microphone for the unknown cases. To demonstrate just how different both recording techniques are, the mean MFCC-vector of the recordings

of the DIA-test for the 249 speakers with known microphone were calculated. Simply visualizing the first three components of this vector results in Figure 6.2. It is clear that classifying between the two microphones is not such a hard job.

To unveil the used microphone of the other samples, an SVM with linear kernel was trained using the R-project for Statistical Computing [98]. To this end, the set of 249 samples with known microphone was split into a training and a test set. Adopting a five-fold cross validation setup on the training set lead to a perfect classification (100%). Training one SVM on the whole training set lead to a performance of 94% on the test set. This last classifier was then used to fill the microphone information gaps in the corpus.

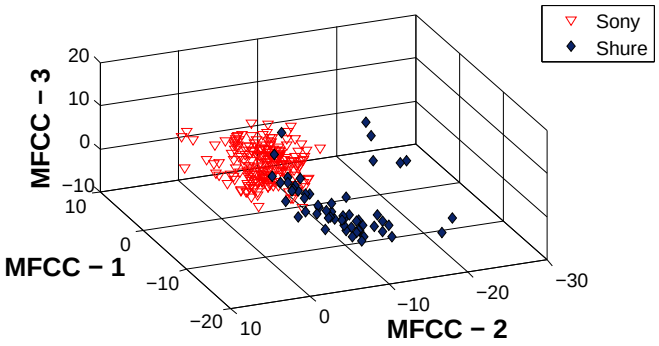

Figure 6.2: First three mean MFCC coefficients of all COPAS speakers indicating differences between the two microphones.

6.3 Spoken Dutch Corpus (CGN)

The CGN (Corpus Gesproken Nederlands, [9, 99, 100]) is a corpus containing Flemish and Dutch speech. It was built between 1998 and 2004 by several Dutch and Flemish universities. It contains about 1000 hours of speech, two thirds originating from the Netherlands and one third from Flanders. It consists of 15 components. Every component contains speech of a different communication setting and acoustic background conditions. The components are summarized in Table 6.2. In this study only the read speech (component-o) was used. All components were orthographically transcribed. For part of the data, phonemic annotations and segmentations were generated taking pronunciation variations into account [100]. These labels and segmentations were generated automatically and manually verified on a word level. Quantitative studies about segmentation procedure and accuracy can be found in [100, 101].

CGN components	
a	Spontaneous conversations (”face-to-face”)
b	Interviews with teachers of Dutch
c	Spontaneous telephone dialogues (recorded via a switchboard)
d	Spontaneous telephone dialogues (recorded on minidisc)
e	Business negotiations
f	Interviews and discussions broadcasted on radio or television
g	Political discussions, debates and meetings
h	Lessons recorded in the classroom
i	Live (a.o. sports) commentaries broadcasted on radio or television
j	News reports and surveys broadcasted on radio or television
k	News broadcasted on radio or television
l	Commentaries, columns and reviews, broadcasted on radio or television
m	Sermons, speeches and ceremonious speeches
n	Lectures and seminars
o	Read speech

Table 6.2: Components distinguished in the Spoken Dutch Corpus.

7

Phoneme intelligibility of monosyllabic words

Using the machine learning foundations layed out in the previous chapters, we are now ready to describe how these fundaments can be used to construct an automated version of the DIA. During this research, several possible approaches were explored. Broadly speaking, they all boil down to a three-stage process involving a front-end analysis, a speaker feature extraction and an intelligibility prediction.

Starting from a speaker's utterance, the **front-end analysis** extracts a stream of acoustic parameter vectors from the waveform. Depending on the speaker feature set, these can be MFCCs (mel-frequency cepstral coefficients [91]) or log-mel-spectral coefficients.

The **speaker feature extraction** considers all these vectors of a speaker to derive a number of global features that characterize this speaker. Deriving interesting speaker features which carry enough information about individual intelligibility problems was one of the main challenges in this research. Originally, all speaker feature extracting methods involved the use of an ASR. As described in Section 5.13, an ASR incorporates a set of reference acoustic models to link a sequence of acoustic feature vectors (MFCCs) to a phone sequence. The creation of these models will be discussed in Section 7.1. Based on these reference models, two strategies were explored for deriving speaker features using an ASR. The first and most straightforward strategy is to just imitate the perceptual test by letting the ASR recognize the targeted phoneme of every word and by measuring the phoneme accuracy (see Section 7.2). A second strategy, involving forced align-

ment, measures how well the speaker's utterance matches the expected phones of the target speech (see Section 7.3).

The **intelligibility prediction model** (IPM) is finally responsible for converting the speaker features into an intelligibility score. Several ML techniques were evaluated to develop robust IPMs. They are described in Section 7.4.

7.1 Reference acoustic models

Reference acoustic models of an ASR are used to describe the probability of a phone being represented by a sequence of acoustic feature vectors $\mathbf{X}$. Traditionally, these models are trained on a large annotated database of speech recorded in similar conditions as in the envisaged recognition task, e.g. all close-talking or far-field speech, with or without noise, etc. Moreover, the type of speech is typically selected to have the same properties as those in the envisaged task, e.g. speakers have the same language and all read similar speech material: running speech or words.

Since we consider non-pathological speech as the reference and as we intend to measure severity of a speech disorder as the degree of deviation from the reference, we developed the reference models using the Flemish non-pathological speech database CoGen. One could argue that this approach establishes that non-pathological speech will be recognized more easily and that the ease of recognizing the uttered speech can be regarded as a measure for intelligibility, as in [3, 102].

However, when confronted with severely disordered speech, the ASR is asked to score sounds that are in many respects very different from the sounds it was trained on. This means that acoustic models are asked to make extrapolations in areas of the acoustic space that were not examined at all during training. One cannot expect that under these circumstances a lower phone probability always points to a larger deviation (distortion) of the observed pronunciation from the norm.

It might therefore be interesting to consider another approach to modeling speech. In [103] the authors describe an acoustic modeling technique employing phonological models to generate an intermediate description of the speech sounds. In this technique, reference acoustic models derive probabilities of a set of binary phonological features given a sequence of acoustical feature vectors. These probabilities are then transformed into a phone probability using a simple product model. Although the reference models are again trained on non-pathological speech, they might offer more potential than the traditional phonetic models typically used for speech recognition when assessing pathological speech degradation. If a speech disorder is localized in specific phonological dimensions, some of the other phonological dimensions of a sound may still be more or less preserved. By deriving phone probabilities from the phonological feature probabilities, there might be a

chance that the ASR can still detect the right phones in the utterance although the probabilities of the phones will be lower than in case of non-pathological speech. In case of phonetic models, it cannot be predicted how the phone probabilities will degrade since acoustic parameter vectors of pathological speech were never examined during training. A second advantage of this phonological detour is that phonological scores may possibly offer a more detailed insight into the articulatory problems behind the speech disorder.

In this research, the potential of two ASRs, further referred to as ASR-ESAT and ASR-ELIS, will be discussed. The next two subsections will dig deeper into their respective acoustic modeling techniques.

7.1.1 State probabilities in ASR-ESAT

ASR-ESAT is a main-stream state-of-the-art ASR [104] developed at ESAT, University of Leuven, Belgium. It comprises three state Semi-Continuous HMMs as the acoustic models. The models represent triphones with tied states. Mixtures of Gaussians from a large set of state independent Gaussians are used as emission distributions. State tying is defined by a global phonetic decision tree with 1567 leaf nodes.

The acoustic front-end derives log-mel-spectral coefficients. At every time $t = 1, .., T$, referring to multiples of 10 ms, a Hamming-windowed segment of 30 ms centered around t is analyzed. Per time step it extracts a vector consisting of 24 Mel spectral coefficients describing the shape of the log-spectrum. To reduce microphone influences, noise masking [105] and spectral mean normalization was performed. By adding the first and second order derivatives of these 24 features, 72 features are obtained. After feature selection using MIDA [106] and decorrelation [107], 39 features X_t are retained.

Based on these parameter vectors, the HMMs compute likelihoods of the form $p(X_t|s_t)$ where $s_t \in \mathcal{S}$ represents a model state at time t. $\mathcal{S}$ denotes the set of all possible states (there are 1567 tied states). The acoustic model scores are converted to posterior probabilities as follows:

$$P(s_t|X_t) = \frac{p(X_t|s_t)P(s_t)}{p(X_t)} \tag{7.1}$$

$$p(X_t) = \sum_{s \in \mathcal{S}} p(X_t|s)P(s) \tag{7.2}$$

The models were trained on the Flemish read-speech part of CGN (component-o) and use a set of 40 Flemish phonemes, being the phonemes which can be found in Appendix A, except for /J/, /E:/, /O:/, /9:/, /Eˆ/, /Aˆ and /Oˆ/, which were mapped on /nj/, /E+/, /A+/, /Y+/, /En/, /An/ and /On/ respectively.

7.1.2 State probabilities for ASR-ELIS

ASR-ELIS was originally developed on American English data by Stouten et al. [103]. We retrained it on Flemish data for this research. As mentioned above, this ASR uses phonological models to create an intermediate description of the speech.

At each time t, which refers to a multiple of 10 ms, the acoustic front-end analyses a Hamming-windowed segment of 25 ms centered around t. Per time step it extracts a vector X_t consisting of 12 MFCCs describing the shape of the log-spectrum and a log-energy describing the total energy of the segment. To reduce microphone influences as described in Chapter 6, Cepstral Mean Subtraction (CMS) was performed.

The phonological models then compute a 24-dimensional vector Y_t for every t. Each component $Y_{ti}, i = 1, ..., 24$ represents the posterior probability $P(A_i|X_{t-5}, .., X_{t+5})$ that phonological class or property A_i $(i = 1, .., 24)$ is "supported by the acoustics" in a 125 ms window around time t, as will be described below. The full list of phonological classes can be found in Table 7.1. The table shows both American English and Flemish phonological classes. The Flemish classes were derived from the standard classification of Flemish phonemes [7, 29]. The Flemish classes differ only slightly from the English ones: /r/ is pronounced retroflex in English while it is pronounced with a trill in Flemish. This explains for instance why retroflex is a class in English but not in Flemish models. Also, no Flemish sounds are classified as dental.

Once the vectors Y_t are determined, they are used to compute posterior probabilities of context-independent phone states at time t. Each phone has exactly one state. The use of context-independent phone states can be justified by the fact that co-articulations between phones are already handled implicitly as the phonological class models make decisions on the basis of a time interval of 125 ms. In order to link the phonological classes to phone states, the canonical values A_{ci} of the phonological classes A_i have to be supplied: $A_{ci} = 1$ means that we desire Y_{ti} to be 1 (on/present), $A_{ci} = 0$ means that we desire Y_{ti} to be 0 (off/absent) and $A_{ci} =?$ means that the class is *irrelevant* for that phone (= both values are equally acceptable). Adhering to the work of Stouten et al. [103], $P(s_t|Y_t)$ is obtained as follows:

$$P(s_t|Y_t) = \left[\prod_{\substack{i=1 \\ A_{ci}(s_t)=1}}^{24} Y_{ti} \right]^{\frac{1}{N_p(s_t)}} \tag{7.3}$$

where $N_p(s_t)$ is the number of phonological classes with a canonical value of 1 for state s_t.

Broad Class	English Subclasses	Flemish Subclasses
voicing	voiced *voiceless* *no activation*	voiced *voiceless* *no activation*
manner of articulation	closure vowel fricative burst nasal approximant lateral silence -	closure vowel fricative burst nasal approximant lateral silence trill
consonant place of articulation	labial labiodental dental alveolar post-alveolar velar glottal	labial labiodental - alveolar post-alveolar velar glottal
vowel place of articulation	low mid-low mid-high high back mid front retroflex rounded	low mid-low mid-high high back mid front - rounded

Table 7.1: Phonological classes for English and Flemish. Italic classes are only used as intermediate outputs and are not displayed as final classification results.

7.1.3 Phonological class probabilities in ASR-ELIS

The phonological features Y_t are computed by a conglomerate of four MLPs, placed in a three-layer architecture. First, the voicing classes (upper row in Table 7.1) are detected. Subsequently, the second layer derives the manner classes (second row in the table). Finally, the third layer consists of 2 MLPs for deriving consonant and vowel place of articulation (third and fourth row in the table respectively). This architecture is depicted in Figure 7.1. All MLPs are fed with the acoustic feature vectors of the incoming speech as well as the outputs of all neural networks higher in the hierarchy. This means that the voicing classes are derived from the MFCCs only, while the other classes also rely on the outputs of MLPs

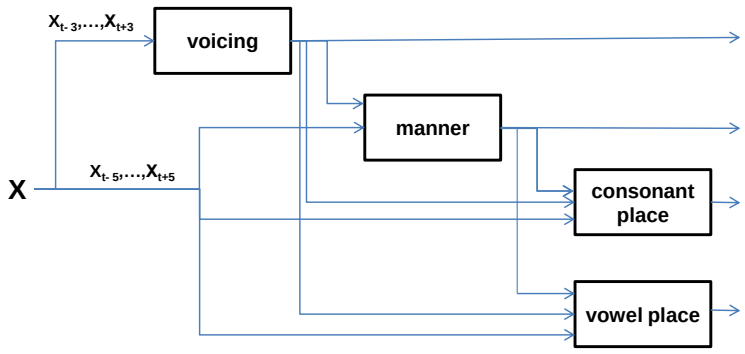

Figure 7.1: Architecture of the phonological feature detector. Following [11].

of previous layers. One can easily understand that manner of articulation benefits from knowledge of the voicing class.

While the voicing network is provided with only three context frames at each side of frame t, the other networks are provided with a context of 5 frames at each side. This extra context can be useful when differentiating between e.g. closure and silence [11].

While the original phonological feature detector (PFD) developed by Stouten et al. was trained on continuous speech uttered by non-pathological American English speakers [103] found in the DARPA TIMIT speech corpus [108], the Flemish PFD was trained on the read speech part of CoGen (see Section 6.1). In a first attempt, a PFD based on the available phone-level segmentation was created, hereafter called CoGeN-1. Also, the voicing and manner networks of the original PFD, based on TIMIT, were reproduced after [11], with one alteration: Cepstral Mean Subtraction was performed to produce equal preprocessing as in case of CoGeN-1. This PFD will be denoted with TIMIT-1.

Classification accuracies of the first two layers (voicing and manner) for both PFDs on independent test sets are shown in Table 7.2 and Table 7.3.

reference class	TIMIT-1	CoGeN-1
voiced	92.4	89.8
voiceless	83.4	69.5
no activation	91.8	89.3

Table 7.2: Classification accuracy (percentage) for the first voicing networks trained on CoGen versus trained on TIMIT. Results on TIMIT are reproduced following [11] but with CMS.

reference feature	TIMIT-1	CoGeN-1
closure	80.0	66.3
vowel	92.7	87.8
fricative	85.8	78.3
burst	70.9	61.9
nasal	80.3	64.9
approximant	51.8	14.7
lateral.	55.2	26.4
trill	-	26.0
silence	91.6	96.8

Table 7.3: Classification accuracy (percentage) for the first manner networks trained on CoGen versus trained on TIMIT. Results on TIMIT are reproduced following [11] but with CMS.

The results clearly show a discrepancy in performance between the networks trained on TIMIT and those trained on CoGeN. The classification accuracies for Flemish are not acceptable as they do not permit an accurate labeling and segmentation by ASR-ELIS anymore. To prove this, a TIMIT-1-based segmenter was used to align the TIMIT-test set and a CoGeN-1-based segmenter was used to align the CoGeN-test set. The resulting segmentations were compared to the corresponding manual phone segmentation delivered with the data. The comparison was made using the Dynamic Time Warping (DTW) procedure described in [11].

As in [11], we mapped the 55 phones in CoGeN to 44 phones and the 58 phones in TIMIT to 48 phones, in order to avoid that we measure substitutions between allophones. Beside phone substitutions we also discern three kinds of segmentation errors: deletions (“del”, a manual segment boundary was omitted), insertions (“ins”, an automatic boundary was inserted between two manual boundaries) and boundary deviations (“far”, the placement of the automatic and the corresponding manual boundary differs by more than 20 ms). The total error is the sum of the segmentation and substitution errors. All errors are specified in percent, relative to the number of phones occurring in the manual labelings. The results in Table 7.4 confirm that the CoGeN-1-based segmenter can not produce a reliable segmentation.

	err	del	ins	far	sub
TIMIT-1	24.2	7.5	6.8	6.3	3.5
CoGeN-1	43.1	13.1	14.4	8.0	7.6

Table 7.4: Evaluation of segmentation and labeling of TIMIT-1- and CoGeN-1-based segmenters. All numbers are percentages.

Since the two PFDs were constructed using the same strategy, the only reason for this discrepancy between English and Flemish results could be the difference in used training database. One main difference is that the TIMIT phone labels and time stamps were manually verified by phonetic experts whereas the CoGeN phone labels were only partly verified and not by experts. Therefore, we wanted to find evidence for the low quality of the CoGeN labels, and if possible, to replace them by more accurate labels. To that end, an HMM-based ASR was trained using the SPRAAK-software [109] to verify these phone labels. Training departed from the orthographic transcriptions in CoGeN, but the given phonetic transcription was supplemented with pronunciation variants as described in [100]. The training was continued until the segmentations generated with the ASR did not change significantly anymore compared to the previous pass. The final segmentations were then used as a reference for the training and evaluation of a second PFD, hereafter called CoGeN-2. Table 7.5 and Table 7.6 show that the new networks for voicing and manner improve significantly, and even do approach the accuracies that were formerly measured on TIMIT.

reference class	TIMIT	CoGeN-1	CoGeN-2
voiced	92.4	89.8	92.3
voiceless	83.4	69.5	85.0
no activation	91.8	89.3	92.9

Table 7.5: Classification accuracy (percentage) for the voicing networks trained on CoGen versus trained on TIMIT. Results on TIMIT are reproduced following [11] but with CMS.

reference feature	TIMIT	CoGeN-1	CoGeN-2
closure	80.0	66.3	67.9
vowel	92.7	87.8	89.4
fricative	85.8	78.3	85.6
burst	70.9	61.9	72.6
nasal	80.3	64.9	80.7
approximant	51.8	14.7	40.3
lateral	55.2	26.4	42.2
trill	-	26.0	52.3
silence	91.6	96.8	96.2

Table 7.6: Classification accuracy (percentage) for the manner networks trained on CoGen versus trained on TIMIT. Results on TIMIT are reproduced following [11] but with CMS.

As a last control, we also measured the segmentation accuracy of the new models in an alignment task. Table 7.7 shows that the trend visible in the phonological classification accuracy is confirmed: the total error rate drops from 43.1%

to 23.9%.

	err	del	ins	far	sub
TIMIT	24.2	7.5	6.8	6.3	3.5
CoGeN-1	43.1	13.1	14.4	8.0	7.6
CoGeN-2	23.9	6.6	9.0	3.4	4.9

Table 7.7: Evaluation of segmentation and labeling of TIMIT- and CoGeN-based PFD-segmenters. All numbers are percentages.

Since the CoGeN-2-PFD-based segmenter attains a similar accuracy as the original TIMIT-based PFD segmenter, the CoGeN-2 networks were used as a base for the Flemish ASR-ELIS.

Now that the acoustic modeling techniques of both ASRs are introduced, we can describe how they were used for appropriate speaker characterization.

7.2 Speaker characteristics based on recognition

A logical first step in the automation of the DIA-test is to let the ASR take the role of the listener and to let it recognize the phonemes. Since the listener can select the right phoneme from a restricted list of possibilities, we also presented only those possibilities to the ASR.

7.2.1 Word Accuracy (WAR)

The simplest speaker feature one can derive from this recognition task is the word accuracy rate (WAR), defined as the percentage of correctly recognized words. A word is considered correctly recognized if the target word obtains the highest score. By computing a WA for the three subtests A, B and C one obtains a set of three WAR-features per ASR. This kind of features is also used in [110].

7.2.2 Log Likelihood Ratio (LLR)

As the WAR is based on a binary decision (word correctly or not correctly recognized), it might be useful to try out a continuous measure to circumvent the effects of discretization. This is done by using the Log Likelihood Ratio (LLR) measure. This measure is defined as the log likelihood of the target (correct) word minus that of the best other word. Here too, three LLR-features can be retrieved per ASR.

7.3 Speaker characteristics based on forced alignment

It is possible to obtain a richer speaker characterization by analyzing the phonetic segmentation made by an ASR which was configured to just align the speech with the target word. There is evidence that measures derived from such an alignment tend to correlate with intelligibility [102].

7.3.1 Phonemic features (PMF)

If the aligner assigns vector X_t to acoustic model state s_t, one can compute the posterior probability $P(s_t|X_t)$. In ASR-ESAT, this requires the conversion of likelihoods $p(X_t|s_t)$ to posteriors according to Equations (7.1) and (7.2). In ASR-ELIS, posterior probabilities $P(A_i|X_{t-5}, ..., X_{t+5})$ are converted to $P(s_t|X_t)$ according to Equation (7.3).

If $F_k(k = 1, ..., N_F)$ is a phone (ASR-ELIS) or phoneme (ASR-ESAT), a phonemic feature PMF(k) for phone(me) F_k can then be derived by taking the mean over the posterior probabilities $P(s_t|X_t)$ of all frames X_t assigned to a state s_t that contributes to phone(me) F_k:

$$\text{PMF}(k) = \langle P(s_t|X_t) \rangle_{t;s_t \in \mathcal{S}_{F_k}} \qquad k = 1, ..., N_F, \tag{7.4}$$

with $\langle a \rangle_t$ denoting the average of a over all t. Repeating this process for every phone(me) gives rise to 40 PMFs for ASR-ESAT and 55 PMFs for ASR-ELIS.

7.3.2 Phonological features (PLF)

Using ASR-ELIS, one can also average the phonological features Y_{ti} (i = 1,...,24). In particular, one can take the mean of Y_{ti} (for some i) over all frames that were assigned to one of the phones characterized by a canonical value $A_{ci} = A$ (either 1 or 0). Such a mean score is thus generally determined by the realizations of multiple phones. Consequently, since different speakers have uttered different word lists, the different phones could have a speaker-dependent weight in the computed means. In order to avoid this, the simple averaging scheme is replaced by the following two-stage procedure:

1. take the mean of Y_{ti} over all frames that were assigned to a phone f having $A_{ci}(f) = A$ (1 or 0), denote this mean as PLF(f, i, A), and repeat the procedure for all valid combinations (f, i, A),

2. compute PLF(i, A) as the mean over f of the PLF(f, i, A) that were obtained in the previous stage.

This procedure gives equal weights to every phone contributing to PLF(i, A). Written in mathematical notation, one gets

$$\text{PLF}(f, i, A) = \langle Y_{ti} \rangle_{t; s_t = f; A_{ci}(f) = A} \qquad \forall \text{ valid } (f, i, A) \quad (7.5)$$

$$\text{PLF}(i, A) = \langle \text{PLF}(f, i, A) \rangle_{f; A_{ci}(f) = A} \qquad i = 1 \,..\, 24; A = 0, 1 \quad (7.6)$$

Since for every of the 24 phonological feature classes there are phones with canonical values 0 and 1 for that class, one always obtains 48 phonological features. The 24 phonological features PLF$(i, 1)$ are called positive features because they measure to what extent a phonological class that was supposed to be present during the realization of certain phones is actually supported by the acoustics observed during these realizations. The 24 phonological features PLF$(i, 0)$ are called negative features. We add this negative PLF set because it is not only important for a patient's intelligibility that phonological features occur at the right time, but also that they are absent when they should be.

In the case of ASR-ESAT, one cannot compute a PLF with the same interpretation, but one can nevertheless introduce the notion of phonological features by adapting the procedure that delivered the PMFs. This is done in 3 steps:

1. Assign the first state of a plosive to a phone of type "closure" (e.g. /#b/, /#p/, /#t/, ...) and the other two to a phone of type "burst"(e.g. /b/, /p/, /t/,...)

2. Compute the PMFs of the new phone set

3. Now calculate the PLF(i,A) as the mean of the PMF(f) over all phones f whose $A_{ci} = A$.

Admittedly, constructing the PLF-ESAT is a bit indirect and gives only an impression of the true phonological features, but nevertheless, using these features we hope to find an answer to the question whether an IPM based on PMFs and PLFs coming from ASR-ESAT can compete with an IPM based on PMFs and PLFs coming from two different recognizers.

Repeating the PLF computation for all phonological features and for two values of A (1 and 0) again results in 48 PLFs per ASR.

7.3.3 Context-dependent phonological features (CD-PLF)

It can be expected that pathological speakers encounter more problems with the realization of a particular phonological class in some contexts than in others. Consequently, it makes sense to compute the mean value of a phonological feature Y_{ti} that takes not only the canonical value of feature class A_i in the tested phone into account but also the properties of the surrounding phones. Since the phonological classes are supposed to refer to different articulatory phenomena, it makes sense to

consider them more or less independently. Due to the ternary nature of the phonological class values (on, off, irrelevant), the number of potential contexts per (i, A) is then limited to 3 x 3 = 9. If we further include "silence" as a special context to indicate that there is no preceding or succeeding phone, the final number of contexts is 16. Taking into account that PLFs are only generated for two canonical values of A, namely 0 and 1 (and not for irrelevant), the total number of context-dependent phonological features (CD-PLF) is 24 x 2 x 16 = 768. This number is however an upper bound since many combinations will not occur in the 50 word utterances of the speaker.

In order to determine in advance all the combinations that are worthwhile to consider in our system, we examined the canonical phonetic transcriptions of the words in the different variants of the A, B or C-list respectively. We derived from these lists how many times they contain a particular combinations. We then retained only those that appeared at least five times in any combination of lists one could make. To determine the number of occurrences of a combination, one just needs to count how many times it occurs in A-list to get an A-count. Similarly one determines a B and a C-count, and one takes the sum of these counts. For our test, we found that 123 of the 768 combinations met the condition we set out.

If A^L and A^R represent the canonical values of feature class A_i in the left and right context phone, the computation of a context-dependent feature for the combination (A, A^L, A^R) is obtained by means of a two-stage scheme:

1. take the mean of Y_{ti} over all frames which were assigned to a phone f having a canonical value $A_{ci}(f) = A$ (A can be either 1 or 0 here) and appearing between phones whose canonical values of class A_i are A^L and A^R, denote this mean as PLF(f, i, A, A^L, A^R) and repeat the procedure for all combinations (f, i, A, A^L, A^R) occurring in the data,

2. compute PLF(i, A, A^L, A^R) as the mean over f of the PLF(f, i, A, A^L, A^R) that were computed in the first stage.

Again, this procedure gives equal weights to all the phones that contribute to a certain CD-PLF. In mathematical notation one obtains

$$\begin{aligned} \text{PLF}(f, i, A, A^L, A^R) &= \langle Y_{ti} \rangle_{t; s_t = f; A_{ci} = A; A^L_{ci} = A^L; A^R_{ci} = A^R} \\ &\qquad \forall \text{ occurring } (f, i, A, A_L, A^R) \qquad (7.7) \\ \text{PLF}(i, A, A^L, A^R) &= \big\langle \text{PLF}(f, i, A, A^L, A^R) \big\rangle_{f;\ \text{occurring } (f, i, A, A^L, A^R)} \\ &\qquad \forall \text{ occurring } (i, A, A^L, A^R) \qquad (7.8) \end{aligned}$$

with A_{ci}, A^L_{ci} and A^R_{ci} being short notations for, respectively, the canonical values of A_i in the state visited at time t, in the state from where this state was reached at some time before t, and in the state which is visited after having left the present state at some time after t.

Note that the context is derived from the phone sequence that was actually realized according to the alignment system. Consequently, if a phone is omitted, a context that was not expected from the canonical transcriptions can occur, and vice versa. Furthermore, there may also be fewer observations than expected for the combination that has the omitted phone in central position. In the case that not enough observations of a particular combination would be available, the corresponding feature is replaced by its expected value (as derived from a set of recorded tests).

7.4 Intelligibility Prediction Models

Once all speaker features are computed, they need to be converted into an objective intelligibility score for the speaker. In doing so we use a regression model that is trained on both the pathological and the normal speakers of COPAS.

7.4.1 Training and evaluation strategies

For these experiments, we used 231_DIA. Since the number of speakers (231) is rather small, we opted for a five-fold cross-validation strategy for model training and validation. Performance is expressed in terms of the Root Mean Squared Error (RMSE) and the Pearson Correlation Coefficient (PCC) between computed and perceptual intelligibilities. The Wilcoxon signed-rank test [111] is used to investigate whether results are significantly different at a confidence level of 0.05.

7.4.2 Experimental setup

A variety of statistical learners is available for optimizing regression problems. However, in order to avoid overfitting, only a few of these can be applied to our data set. We used Support Vector Regression (SVR), Random Forests (RF) and Ensemble Linear Regression (ELR), all discussed in Chapter 5. The last two are used in combination with feature selection.

During training of all models, we used RMSE and not PCC as optimization criterion. The reason for this is that we wanted the computed scores to approximate the correct scores directly while PCC actually quantifies the degree of correlation between the correct scores and the best linear transformation of the computed scores. Another reason for using RMSE is that it is directly interpretable. In case the discrepancies (errors) are normally distributed, 67% of the computed scores lie closer than the RMSE to the measured (correct) scores. Anticipating on the next paragraphs, we used the Lilliefors test [111] to verify that in almost all experiments the errors were indeed normally distributed.

7.4.2.1 Ensemble Linear Regression

The ELR technique was implemented in C using the GNU Scientific Library [112] for multilinear regression purposes. We implemented bagging strategies and feature selection processes. Per training fold, ten linear regression models incorporating feature selection are created. These models are then combined into one single model by just averaging the regression coefficients of the submodels. The final model is then evaluated on the validation fold. In a five-fold cross-validation setup this leads to 50 submodels, all created with feature selection. The most important features will appear in many submodels, leading to a feature ranking from which potentially important relations between features and predictions can be derived.

7.4.2.2 Support Vector Regression

The SVR is achieved by a Support Vector Machine (SVM) built with libsvm [113] with a Gaussian or a linear kernel. During the training of the SVR on a particular training-validation partition, we select the learning parameters (kernel parameters, fault threshold) by means of a grid search based on an internal 5-fold cross validation within the training part of the partition.

7.4.2.3 Random Forests

Random Forests were also evaluated for regression and feature selection. Using a C++ implementation of Random Forests, we performed feature selection for regression, and used the most important features to build a final regression model.

7.4.2.4 Evaluation of results

For every single feature set, three IPMs were evaluated: one using ELR, one for RF and one for SVR. All IPMs are trained and evaluated using a 5-fold cross validation (CV) strategy.

All statistical learners are programmed to optimize the RMSE. Validation results are summarized in Table 7.8. Per feature set, only the best performing learner is displayed.

The Wilcoxon signed-rank test [111] revealed that differences between the best results in this table (marked in bold) and all others are significant at a confidence level of 0.05.

A first striking result is the fact that ELR delivers the best results for most IPMs. ELR is thus a good choice for this regression problem: it does not only achieve the best performance but it is also way simpler and faster to create. While the results for RF are way worse, SVR attains almost the same accuracy as ELR, using sometimes a RBF and sometimes a linear kernel. The relationship between

feature set	validation results		best learner
	RMSE	PCC	
WAR-ELIS	12.44	0.33	ELR
WAR-ESAT	**9.04**	**0.73**	SVR-RBF
LLR-ELIS	12.57	0.24	ELR
LLR-ESAT	11.98	0.46	SVR-RBF
PMF-ELIS	11.50	0.58	ELR
PMF-ESAT	9.24	0.70	ELR
PLF-ELIS	**8.20**	**0.78**	ELR
PLF-ESAT	9.56	0.69	ELR
CD-PLF-ELIS	9.44	0.72	SVR-LIN

Table 7.8: RMSE and PCC between computed and perceptual intelligibility scores for all individual IPMs represented by the utilized feature set and the ASR system from which it emerges. The last column indicates which statistical learner yielded the best validation results. Results above the horizontal line originate from recognition-based feature sets, while those under the line originate from alignment-based feature sets. The best result under and above the line are marked in bold.

most feature sets and intelligibility score seems to be linear. Only WAR-ESAT benefits from a RBF kernel.

If we now take a closer look at the results of the individual IPMs, we can notice three important things. First of all, feature sets originating from recognition tasks, i.e. simply repeating the perceptual task, correlate less well with intelligibility scores than the feature sets derived from forced alignment. Furthermore, feature sets deduced in a more natural way perform better than those deduced with a deviation: PLFs derived from ASR-ELIS outperform those derived from ASR-ESAT while PMFs derived from ASR-ESAT strongly outperform those derived from ASR-ELIS. Notice that CD-PLFs, derived from ASR-ELIS, perform worse than PLF-ELIS features. This might be because they extract too detailed information, leading to overfitting to the training part in the submodels. The individual feature set with the most predictive power is PLF-ELIS with a RMSE of 8.20 between computed and perceptual scores.

An interesting issue is the nature and the number of selected features. Using ELR one can easily derive this information. Since every training fold was based on 10 submodels, a total of 50 small linear regression models was built, each of which was free to choose its own features. Having access to these 50 models, we calculated the mean number of selected features as well as the popularity of each feature. The latter is defined as the number of submodels in which a feature was selected. Table 7.9 shows the mean number of features selected by a submodel.

For the recognition-based feature sets, which only have a cardinality of 3, all features are selected in every model. Focusing on the two most predictive

feature set	Ns
WAR-ELIS	3
WAR-ESAT	3
LLR-ELIS	3
LLR-ESAT	3
PMF-ELIS	3
PMF-ESAT	4
PLF-ELIS	7
PLF-ESAT	4
CD-PLF-ELIS	4

Table 7.9: Mean number of selected features per random subfold in ELR.

alignment-based sets PLF-ELIS and PMF-ESAT, the 10 most popular features for each feature sets are listed in Table 7.10.

PLF-ELIS		PMF-ESAT	
selected features	N	selected features	N
not front	50	/i/	50
fricative	48	/A+/	38
silence	48	/s/	27
voiced	36	/z/	23
high	32	/p/	23
front	27	/2/	22
mid-low	17	/d/	20
no silence	15	/w/	16
burst	14	/k/	11
mid	12	/l/	10

Table 7.10: Most selected features and the number of submodels N in which they were used.

An important question is whether the selection of features by the two different IPMs can be motivated from a clinical point of view. When analyzing Table 7.10, seven groups of related features can be distinguished: (1) vowel-related features, (2) silence-related features, (3) fricative-related features, (4) voicing-related features, (5) plosive-related features, (6) lateral-related features and (7) approximant-related features. It is clear that the most important features are related to vowels. Half of the most predictive PLFs and 3 of the most predictive PMFs refer to vowels. In classic speech theory about the perception of normal speech, consonants were seen as the primary information-bearing elements of speech [114]. However, more recent research suggests that vowels make an important contribution to the intelligibility of normal as well as pathological speech. Persons with reduced

tongue motility are known to show a strong correlation between intelligibility and vowel trapezium size [115]. Ansel and Kent [116] found that word intelligibility in dysarthria associated with mixed cerebral palsy can be predicted with 62% accuracy by three vowel-related phonological contrasts and only one consonant contrast. The three vowel-related contrasts involved were: front-back, high-low and tense-lax. A study of Kent et al. [117] revealed a strong correlation between tongue height and word intelligibility in speakers with dysarthria secondary to ALS.

Also for hearing impaired persons it is known that the formant frequencies of vowels tend to converge toward those of the central vowel /@/. The reduced phonological space is most present in the vowel place dimension. Front-back tongue movement is not very visible from the outside and is thus not simple to mimic [118]. Liu et al. also demonstrated that esophageal speakers had significantly higher formant frequencies (F1, F2, and F3) and a significantly diminished vowel space area compared to laryngeal speakers [119]. The correlation between accurate vowel production and intelligibility has been stressed by two studies on the vowel space area. The vowel space area, constructed from F1 (related to jaw opening and tongue height) and F2 (related to place of constriction or horizontal tongue position), is defined as the area of the trapezium connecting the corner vowel representations in the F1-F2 space [114]. In many speech disorders, articulatory displacements are often reduced, resulting in a compressed vowel space area. Liu et al. [119] found a significant correlation between vowel space and intelligibility at vowel (0.63) and word level (0.68). Consequently, the high prevalence of vowel-related features in the IPMs is in agreement with these studies.

Two of the ten most important phonemic features are fricative related. Actually, two of the six Dutch fricatives are selected by the PMF intelligibility model. The phonological feature “fricative” is also one of the features in the PLF model. Previously, acoustic features of the fricative /s/ were shown to highly correlate with dysarthric speaker’s overall intelligibility [120]. Additionally, three of the 19 phonetic contrasts assessed by the intelligibility test of Kent et al. [121] are related to fricatives. Laryngectomees are also known to encounter problems with the voicing of fricatives and plosives [47, 55], which also explains why the voiceless-voiced-pair /s/ - /z/ occurs in the list of most important PMFs and both ‘fricative’ and ‘voiced’ are amongst the most important PLFs.

The intelligibility assessment of Kent et al. also indicates the importance of an adequate production of plosives. The importance of plosives is supported by the occurrence of /p/, /d/ and /k/ in the most important PMFs and the PLF ‘burst’ occurring in 14 of the 50 submodels. Other pathologies also support the presence of plosive, fricative and voicing related features: persons with cleft lip and palate lose part of their mouth pressure through the nose when realizing voiceless fricatives

and plosives [38] and for hearing impaired persons most errors occur in phones produced in the middle or the back of the mouth, like e.g. palatal plosives and fricative sounds. One of the most frequent errors in consonant production is the voiced-voiceless confusion, most often voiced for voiceless substitution [62].

Silence was also found to be an important feature. A possible explanation for this is that pathological speakers often make extraneous sounds before and after an utterance or during closures. In this respect, the selection of silence could be related to the selection of plosives.

Although the phonemic features /l/ and /w/ are seemingly necessary to achieve a strong correlation between the perceptual and objective intelligibility measures, no substantial evidence for the importance of these features could be found in the literature. The fact that the contrasts /r/ - /l/ and /r/ - /w/ are included in the single-word intelligibility test of Kent et al. [121] can however be interpreted as an implicit acknowledgement. The test examines 19 phonetic contrasts which were selected to present the speech problems experienced at the segmental level by speakers with dysarthria.

7.4.3 Combination of feature sets

From the former experiments, it follows that none of the individual speaker feature sets leads to a correlation between the objective and the perceptual scores that can compete with the inter-rater agreement of 91% observed between individual raters. In this respect, we have investigated whether the combination of different feature sets may bridge this gap. Since Table 7.8 showed that WAR-ESAT and PLF-ELIS were the most predictive recognition-based and alignment-based feature set respectively, we only tried out combinations that contained at least one of those two, and we only added feature sets that individually lead to a RMSE below 10, i.e. WAR-ESAT, PLF-ELIS, PLF-ESAT, PMF-ESAT and CD-PLF(-ELIS).

There are several ways of constructing an IPM that combines two feature sets. In this work, we explored two strategies: early fusion and late fusion. Both are displayed in Figure 7.2. Early fusion (IPM 3 in the figure) combines both feature sets into one set, which is subsequently supplied to the IPM. Late fusion (IPM 4 in the figure) uses the outputs of two IPMs trained on individual feature sets and combines the results of these two models to obtain the final result. The combination of two IPMs can again be accomplished in several ways, and can be as complex as performing an extra SVR to map both individual model outputs to the intelligibility score. However, this would require an extra cross-validation loop. Here, we simply calculated the final intelligibility score as the mean of the individual intelligibility scores.

Since RF was outperformed by far by the other learners in previous experiments, early and late fusion were only combined with SVR- (linear and RBF

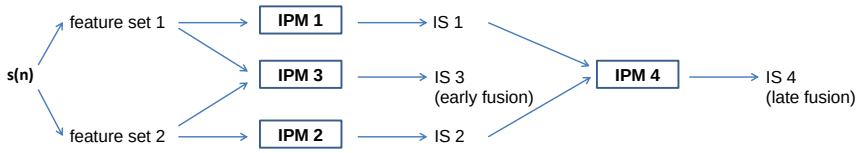

Figure 7.2: Early and late fusion. Predictions of the Intelligibility Scores (IS)

kernel) and ELR-based IPMs. Results are shown in Table 7.11. For each learner, only the best technique (early or late fusion) is displayed.

feature set	ELR results			SVR results			
	RMS	PCC	fus	RMS	PCC	kernel	fus
PLF-ELIS + PMF-ESAT	7.91	0.79	late	7.96	0.80	RBF	early
PLF-ELIS + PLF-ESAT	8.20	0.78	early	8.31	0.78	LIN	early
PMF-ESAT + WAR-ESAT	7.96	0.80	late	7.64	0.82	RBF+RBF	late
PLF-ELIS + WAR-ESAT	**7.86**	**0.80**	**early**	**7.44**	**0.82**	RBF	**early**
CD-PLF + PLF-ELIS	8.19	0.78	early	8.18	0.78	RBF	early
PLF-ELIS + PMF-ESAT + WAR-ESAT	8.00	0.79	early	7.66	0.81	RBF	early

Table 7.11: RMSE and PCC between computed and perceptual intelligibility scores for IPMs supplied with two feature sets. Per learner, the best result is put in bold. It is also used as a reference for significance tests. Underlined results differ significantly from the reference. The penultimate column indicates which kernel yielded best SVR results. In case of late fusion, kernels are mentioned per feature set (in the same order as the first column). RMS denotes RMSE, fus denotes fusion.

The results show that combining feature sets does generally not improve the results. The Wilcoxon signed-rank test reveals that only the combination of the individually best performing feature sets PLF-ELIS and WAR-ESAT leads to a significant improvement at a confidence level of 0.05 over PLF-ELIS. The combination of PLF-ELIS and WAR-ESAT leads to a RMSE as low as 7.44 and a PCC of 0.82 using a SVR with a RBF kernel. Note that the best results for SVR and those using ELR are not significantly different. Adding more feature sets to this best combination did not further improve the performance. A possible explanation

for this is that the simple forward feature selection technique of ELR is inadequate for searching the optimal feature combination in a high-dimensional feature space. SVR on the other hand might lose generalization power in higher feature space dimensions.

Given the performances in Table 7.11, we retain two feature set combinations. First of all, WAR+PLF-ELIS because it reaches the best results, and secondly, PLF-ELIS + PMF-ESAT because its results are not significantly worse than the best results. Moreover, PMFs can provide more insight in important underlying dimensions of intelligibility than WAR. Also, we retain ELR and SVR with a RBF kernel as learners since SVR with a linear kernel is only selected once.

In order to assess the accuracy of our best system (PCC = 0.82) in relation to human listeners' performance (ICC = 0.91), we have to take into account that PCC and ICC are different measures and that the reported ICC was obtained on a set of only 30 samples. Since we do not have access to the data leading to this ICC, we therefore cannot determine whether these 30 samples were representative or not. Anticipating on the next paragraphs, we can easily select a group of about 30 samples (the hearing impaired speakers), out of the 231, for whom the current model leads to a RMSE of 7.72 and a PCC of 0.94. This illustrates that a PCC based on 30 samples does not necessarily prove a statement for a group of 231 samples.

In this respect, a speech therapy student in her final year recorded 61 new speakers of different pathologies, reading the DIA test. These 61 samples were judged by 2 speech therapists: the student and the experienced speech therapist who judged the samples of COPAS. Having access to both ratings, a RMSE between the two of 7.0 and a PCC of 0.89 was determined [122]. Comparing this to our best RMSE of 7.44, the latter experiment supports the conclusion that our automatic system exhibits a close-to-human accuracy.

7.4.4 Pathology-specific models

If a clinician is mainly working with one pathology, he or she is probably more interested in an intelligibility prediction model that is specialized to that pathology. As a starting hypothesis, we assume that - since people with different pathologies are bound to have different articulation problems - pathology-specific models can differ from general models.

7.4.4.1 Establishing a baseline per pathology

Before exploring the pathology-specific models, we first investigate the prediction results of the best general models (based on WAR+PLF-ELIS, hereafter called

WAR+PLF, and PLF-ELIS+PMF-ESAT, hereafter called PLF+PMF) on the pathology-specific samples. Since we did not record the validation performances per pathology, we had to repeat the five-fold cross-validation experiments of these models as before, but now with a measurement of the pathology-specific accuracies.

The performances of the general models for dysarthria (D), laryngectomy (L), hearing impairment (H) and cleft (C) can be found in Table 7.12. We only evaluated early fusion because this was almost always the best strategy in Table 7.11. Since the pathology-specific groups are rather small, we tested significance of differences at the level of 0.05 and 0.10. If larger groups would be available, it would maybe be possible to prove significances with more confidence. However, for the moment this is only a hypothesis which needs experimental verification.

P	N	feature set	ELR results		SVR results		
			RMSE	PCC	RMSE	PCC	kernel
D	74	**WAR + PLF**	8.57	0.78	**7.90**	**0.83**	**RBF**
		PLF + PMF	8.97	0.74	8.72	0.78	RBF
H	29	**WAR + PLF**	**6.33**	**0.92**	6.66	0.89	RBF
		PLF + PMF	6.46	0.97	6.66	0.89	RBF
L	30	**WAR + PLF**	8.95	0.73	**8.87**	**0.73**	RBF
		PLF + PMF	9.92*	0.72*	10.89	0.61	**RBF**
C	38	**WAR + PLF**	7.73*	0.43*	**6.48**	**0.63**	RBF
		PLF + PMF	7.60*	0.38*	6.66	0.55	RBF

Table 7.12: RMSE and PCC between computed and perceptual intelligibility scores for the most predictive general IPMs. The first column "P" denotes the pathology: D(ysarthria), H(earing impairment), L(aryngectomy) or C(left). The second column denotes the number (N) of persons per pathology. The last column indicates which kernel yielded the best SVR results. Per pathology, the best results are indicated in bold, and results differing significantly with $p \leq 0.05$ are underlined, with $0.05 < p \leq 0.10$ are denoted with $*$.

A first result is that the best general model (IPM with WAR + PLF using SVR-RBF) remains the best model for all pathologies, be it that for hearing impaired and cleft speakers, the models based on PLF + PMF are not significantly worse. Concerning the used learner, we can conclude that only ELR and SVR-RBF yield satisfactory results.

Table 7.12 also illustrates that PCC and RMSE results do not always point in the same direction. Take for example the results for the hearing-impaired and for clefts, more specifically the results for an IPM based on PLF+PMF using an SVR with a RBF kernel. Scatter plots of the total speaker group, clefts and of the hearing impaired speakers are presented in Figure 7.3. While the RMSEs are exactly the same, the PCCs differ more than 30% absolute. The difference lies in the range

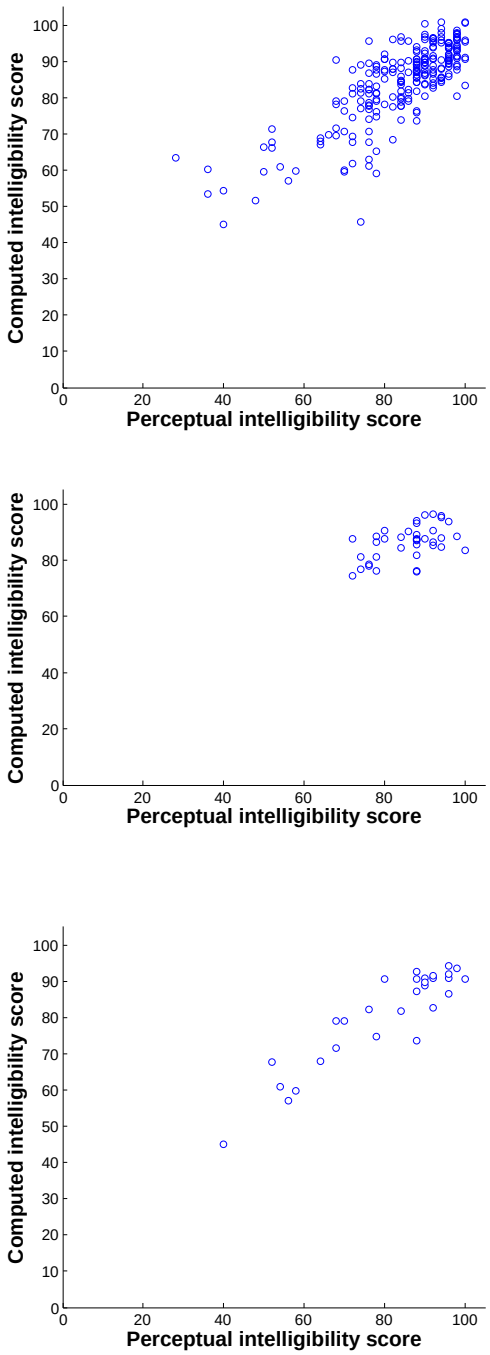

Figure 7.3: Results of the general model using PLF+PMF with SVR-RBF for all (top), cleft lip and palate (middle) and hearing-impaired speakers (bottom).

of the intelligibility scores. If a model (the general model in this case) is designed to cover a large intelligibility range (28-100), and if it is evaluated on a subgroup covering only a small subrange (clefts: 72-100, hearing impaired: 40-100), the PCC can be quite low for this subgroup even though the errors remain acceptable. This happens when the rankings of the speakers of this group are significantly different along the perceptual and the objective scores respectively.

7.4.4.2 Pathology-adapted models

Having evaluated a baseline IPM, we can now construct pathology-specific IPMs. The coefficients of the baseline IPM have been trained on all training samples and validated on the pathology-specific samples of the validation fold. If we denote the totality of the samples as T, then T consists of two parts: T_t and T_v, being the training and the validation part respectively. In this section, we will use the notations P_t and P_v for the pathology-specific samples of T_t and T_v respectively. Note that $P_t \in T_t$ and $P_v \in T_v$ and that P_t and P_v together constitute P.

Nothing inside the baseline IPM was tuned to P_t. In an attempt to create pathology-specific IPMs, we investigated how we could add P_t-related information to the model.

The most simple approach would be to train and test a model on one pathology only. However, we may have too few speakers per pathology (29-74) to compute reliable regression coefficients of such a model. Therefore, we explored the Domain Adaptation techniques offered by Daume et al. [89], as explained in Section 5.12. Per pathology, per feature set combination (WAR+PLF or PLF+PMF) and per learner (ELR or SVR-RBF), we created 6 IPMs, all using the same five-fold division:

- baseline IPM (B), same as in Table 7.12: trained on T_t and validated on P_v,
- weighted samples (WS): trained on $T_t + (w-1).P_t$ and validated on P_v, with w an external parameter denoting the weight given to the samples of P_t (note that training on $T_t + (w-1).P_t$ gives weight w to P_t since T_t already contains P_t once),
- weighted models (WM): Model 1 (M_1) trained on T_t, Model 2 (M_2) trained on P_t, $(M_1 + w.M_2)/(1+w)$ validated on P_v, with w again an external parameter denoting the weight.
- cascade model (C): Model 1 trained on T_t, resulting intelligibility scores are used as an extra feature for Model 2 trained on P_t and validated on P_v.
- feature doubling (FD): Samples of P with feature vector X now obtain feature vector $[X, 0, X]$, other samples of T obtain feature vector $[X, X, 0]$.

With these new feature vectors, a model is trained on T_t and validated on P_v.

- P only (PO): trained on P_t and validated on P_v.

The extra weight w for WS was chosen in such a way that w times the cardinality of P_t is about as high as the cardinality of T_t minus the cardinality of P_t so as to give equal importance to the two parts in the cost to minimize (so $w = 7$ for hearing impairment and laryngectomy, 5 for cleft and 2 for dysarthria). Table 7.13 reports the results for these 6 IPMs in terms of RMSE. It shows that domain adaptation techniques can sometimes improve the baseline results (a significance level of 0.10), but none of the adaptation techniques is consistently the best. This suggests that the few improvements which are achieved by using domain adaptation might be rather chance hits since no conclusions on which adaptation method to use can be drawn from these results. More data will help to clarify this point.

Table 7.13 also shows that the results of domain adaptation on ELR-based models often outperform those of SVR-based models. This might be because with SVR we are creating too complicated models for only a small dataset, while the feature selection in ELR automatically corrects for that.

As expected, models PO trained only on one pathology do not improve upon the baseline, but usually perform a lot worse. This is probably due to the low cardinality of the number of samples for specific pathologies.

Since the general model for hearing-impaired speakers built with ELR and based on PLF + PMF was significantly worse than the domain-adaptation model (weighted samples, ELR, PLF + PMF), we investigated whether different features were selected for the two models. Features appearing more than 10 times in the general and in the specific model are listed in Table 7.14. Apparently there are quite some differences between both IPMs. One vowel-related features was added, reinforcing the importance of vowels. The addition of the feature 'nasal' confirms the findings of [63, 123, 124] that hearing impaired speakers have trouble finding the right nasality. Addition of the plosive /G/ complies with the fact that phonemes produced in the middle or the back of the mouth are more prone to errors [62]. The features 'not fricative' and 'not closure' could also be related to these visibility issues.

7.5 Conclusions for this chapter

In this chapter, we described our first steps toward an automation of the perceptual DIA test. Basically, our methodology consists of three steps. Starting from a speaker's utterance, the **front-end analysis** extracts a stream of acoustic parameter vectors from the waveform. The **speaker feature extraction** considers all these vectors of a speaker to derive a number of global features that characterize

P	method	WAR + PLF		PLF + PMF	
		ELR	SVR	ELR	SVR
D	B	8.57	7.90*	8.97	8.72
	WS	8.16*	**7.45**	9.35	8.50
	WM	9.19	9.93	9.07	9.26
	C	8.37*	9.39	8.81	9.22
	FD	8.73	9.82	8.92	11.98
	PO	9.25	9.93	9.20	9.26
H	B	6.33*	6.66*	6.46*	6.66*
	WS	5.96	7.46	**5.15**	6.35*
	WM	6.88*	9.51	6.76*	7.72
	C	6.32*	9.72	6.94*	7.94
	FD	6.56*	7.70	6.98*	10.79
	PO	7.62	9.51	7.94	7.72
L	B	8.95	**8.87**	9.92*	10.89
	WS	9.44	9.48	10.98	10.92
	WM	9.69*	11.25	9.41	11.86
	C	9.26	11.42	10.77	11.74
	FD	8.97	9.79*	10.49	12.21
	PO	10.39	11.25	9.99*	11.86
C	B	7.73*	**6.48**	7.60*	6.66
	WS	7.82*	7.55*	7.33*	6.68
	WM	6.76	8.05*	7.20*	8.52
	C	6.95	8.00*	6.96	8.61
	FD	7.32*	7.23*	6.96	7.17*
	PO	7.28*	8.05*	7.51*	8.52

Table 7.13: RMSE between computed and perceptual intelligibility scores after domain adaptation of IPMs built with PLF+PMF and WAR+PLF. Per pathology, the best result is marked in bold. Results differing significantly from the best result ($p > 0.05$) are underlined. If $0.05 < p \leq 0.10$, the result is denoted with $*$.

this speaker. The **intelligibility prediction model** (IPM) is finally responsible for converting the speaker features into an intelligibility score.

For the speaker feature extraction, two strategies were explored. The first and most straightforward strategy was just imitating the perceptual test by letting the ASR recognize the targeted phoneme of every word and by measuring the **word accuracy rate** (WAR). A second and novel approach was an alignment-based approach. Starting from a state-of-the-art ASR, using the traditional phonetic models for Flemish (normal) speech, the pathological speaker's utterance was aligned with its target phonemic transcription, from which a number of global features that were derived. The latter constitute the **phonemic characterization of the speaker**, describing how well on average the acoustic realizations of the phonemes are scored

general	hearing-impaired
voiced	voiced
fricative	fricative
burst	burst
-	**nasal**
silence	silence
mid	mid
front	front
rounded	rounded
-	**not closure**
-	**not fricative**
-	**not mid**
not front	not front
/d/	/d/
/k/	/k/
/s/	/s/
/z/	/z/
-	**G**
/m/	/m/
/i/	/i/
/A+/	/A+/

Table 7.14: Features selected by the general IPM and the hearing-specific IPM based on WAR+PLF+PMF. Features in bold indicate differences between both models.

by the phonetic models of the ASR. Since speech intelligibility is closely related to articulation, we investigated whether phonological models of speech could offer more potential than the traditional phonetic models. Starting from an ASR which uses phonological models for (normal) Flemish speech, the pathological speaker's utterance was therefore aligned with its phonetic transcription, from which a number of global features were derived that constitute the **phonological characterization of the speaker**. The latter describe how well on average the phonological realizations of the speaker are scored by the ASR.

The potential of the three feature sets was then explored by building several IPMs, revealing that a combination of WAR and phonological features leads to intelligibility scores that correlate very well (Pearson Correlation Coefficient higher than 0.80) with the perceptual ratings. Furthermore, using both alignment-based feature sets, one achieves correlations which are not significantly lower than the best results. Combining more than two feature sets did not lead to better results.

Departing from the general IPMs, domain adaptation techniques were explored to build pathology-specific models. We showed that domain adaptation techniques can slightly improve the general model's results for hearing impaired and

dysarthric speakers, leading to correlations up to 0.96 between perceptual and automatic intelligibility scores, but for laryngectomees and persons with cleft lip or palate they provide little to no improvement.

The correlations for general and pathology-specific models compete with the inter-rater agreements measured for perceptual intelligibility assessment. We can therefore conclude that phoneme intelligibility can be predicted in a reliable way by our methodology.

By investigating which features were important in automatic intelligibility prediction, we found that all features frequently selected by the intelligibility prediction models can be linked to known articulatory deficits of pathological speakers. This opens the door for a more profound characterization of pathological speech.

8 Running speech intelligibility

8.1 Introduction: why running speech?

In the previous chapter, we showed that an ASR can be used as an objective listener for determining intelligibility. We showed that it is possible in this way to automate the DIA and to obtain a reliable phoneme intelligibility score in this way.

There are however some minor drawbacks when using this test. A first problem with the test is that phoneme intelligibility (PI) is only moderately correlated with the ability to communicate in more realistic situations where running speech is used [68, 121]. A second problem is that especially children tend to misread a nonsense target word as a more common existing word. These errors obviously induce a negative bias in the speaker's intelligibility. Because of these problems, we envision an automated test that utilizes running speech and that is robust against reading errors, hesitations, etc. of the speaker.

8.2 Predicting running speech intelligibility for COPAS

In order to develop and evaluate the envisaged running speech intelligibility (RSI) models, we first conducted experiments with a part of COPAS, in particular the subset of 121 speakers which have read the paragraph (TM) as well as the DIA. As described in Chapter 6, 121_TM and 121_DIA consist of 47 dysarthric, 26 hearing impaired, 15 laryngectomized, 6 voice disordered, 1 glossectomized and 26 con-

trol speakers. Perceptual PI scores (derived from the DIA recordings) are available for all speakers, but no RSI scores.

Relying on the known (moderate) correlation between PI and RSI [2], we contemplate that - if speaker features computed on running speech can be converted to PI - they can be converted to RSI as well, provided that perceptual RSI scores were available for model training.

Starting from the running speech recordings, we applied the same three-stage process strategy as in Chapter 7, consisting of a front-end analysis, a speaker feature extraction and an intelligibility prediction.

The front-end analysis could be extended towards running speech without any changes.

Concerning the feature extraction part, two feature set combinations led to reliable phoneme intelligibility scores in Chapter 7: WAR + PLF and PLF + PMF. Although WAR was a very predictive feature set for 231_DIA, it cannot be employed for running speech in a straightforward way since it requires a continuous speech recognizer which is much more complex than the isolated word recognizer we used before, especially if we would like to develop a method that is also text and language independent (see later). On the other hand, the alignment-based feature sets can be extended towards running speech without any change of concept. We therefore only considered the feature sets PLF, PMF and their combination.

The process of creating IPMs could remain unchanged. To construct a baseline, we first created IPMs based on PLF, PMF and PLF+PMF derived from 121_DIA, which is just a subset of the samples used in Chapter 7. Then, new IPMs were created based on PLF, PMF and PLF+PMF, but this time derived from the running speech set 121_TM (same speakers uttering the paragraph). However, since no RSI scores were available, we used the PI scores (targets of 121_DIA) as a proxy for the RSI scores. Again, we developed two models per feature set (combination): one based on ELR and one based on SVR with a RBF kernel.

Table 8.1 presents the results of these first PI predictions based on running speech (right), compared to the results we obtained for the same speaker set by employing the DIA recordings as train and validation data (left).

Firstly, the results on 121_DIA seem to be consistent with the results in Chapter 7. The PLFs are again the most predictive features, leading to an average RMSE of 7.71 between computed and perceptual PI scores. Adding other features does not improve upon the results. The results on 121_TM confirm these conclusions but the RMSE's are slightly higher (see below).

Secondly, the results show that the PLFs perform worse on running speech than they do on isolated words, while the PMFs perform equally well. Possible phenomena which play a role in these differences are: (1) the used target scores and (2) the used test material. Since we used PI scores based on DIA instead of RSI

feature set	results on 121_DIA			results on 121_TM		
	RMSE	PCC	learner	RMSE	PCC	learner
PLF	**7.71**	**0.81**	**ELR**	8.80	0.75	ELR
PMF	9.33	0.71	ELR	9.20	0.72	ELR
PLF + PMF	8.18	0.80	ELR-late	**8.57**	**0.77**	ELR-late

Table 8.1: RMSE and PCC between computed and perceptual PI scores starting from DIA (left) and RSI recordings (right). Columns 4 and 7 show the used learner for these results. ELR-late denotes late fusion of two IPMS built with one feature set. Results in the same column differing significantly from the best result at a level of $p < 0.05$ (indicated in bold) are underlined.

scores, a drop in performance could have been anticipated for all models trained on 121_TM since there is no 1-to-1 relation between PI and RSI. On the other hand, since the acoustic models of both ASRs were trained on running speech, the new test material of 121_TM offers contexts which match better with those seen during training than the those seen in isolated (nonsense) words of 121_DIA. Especially for ASR-ESAT, which employs context-dependent models, running speech models can be expected to perform much better on running speech than on isolated words, leading to higher posterior phoneme probabilities on running speech. Apparently, for PMFs, the negative impact of the former effect is compensated by the positive impact of the latter effect, while this is not the case for PLFs. Another effect of using running speech material is that it takes forced alignment at a sentence level. We argue that a forced alignment of isolated words is less prone to errors than forced alignment of sentences, especially in case of pathological speech. This could also cause a drop in performance for all IPMs trained on 121_TM.

Note that the RMSE of the best RSI-based IPM is still around 10% higher than the best RMSE using DIA recordings. In this respect it is interesting to mention that the differences between the DIA-based PLF results and the TM-based PLF+PMF results are not statistically significant though.

At present, our system already achieves a very good prediction of phoneme intelligibility starting from running speech, proving that it is possible to use more natural text material than the 50 partially nonsense-words to obtain a reliable PI score. We anticipate that our system will thus be able to predict RSI as well.

8.3 Towards an alignment-free characterization of speech

Although we have made it plausible that RSI can be predicted with alignment-based techniques, we would like to point out some limitations to this approach. As already mentioned, especially children tend to make reading errors. These errors obviously induce a negative bias in the speaker's intelligibility.

We contemplate that reading errors and hesitations can give rise to alignment errors (in alignment-based methods) or recognition errors due to Out-Of-Vocabulary (OOV) words (in recognition-based methods). This is already the case for reading errors made by normally speaking children [125], let alone for errors that are made by children with a speech disorder where reading skills are often less developed. Because of these problems, we envision an automated test based on running speech that is robust against reading errors, hesitations, etc. of the speaker. This is why we propose a novel approach not involving any alignment or recognition.

A first attempt to predict speech intelligibility without alignment was made by Bocklet et al. [126]. In that work, a speaker verification approach was adopted: a GMM was trained for every speaker, and the parameters of that GMM were used as features from which to predict the speaker's intelligibility.

The approach proposed here relies on a phonological analysis of the uttered speech using phonological feature detectors (PFDs). Two strategies were developed:

- In a first strategy, every output of the PFDs is statistically analyzed separately, leading to a phonological feature representation of the utterance, called alignment-free phonological features (ALF-PLFs). This strategy relies on the hypothesis that intelligibility reduction due to a certain speech disorder could be owed to problems with the realization of individual phonological classes.

- A second strategy uses the outputs of the PFDs to calculate posterior phone probabilities which in turn are statistically analyzed, leading to a phonetic representation of the utterance, called alignment-free phonetic features (ALF-PMFs). This strategy relies on the hypothesis that intelligibility degradation is correlated with problems in realizing a certain combination of phonological classes as needed for producing the phones.

8.3.1 Alignment-free phonological features

Like for the other IPMs, derivation of RSI scores is a three-stage process involving a front-end analysis, a speaker feature extraction and an intelligibility prediction. The front-end analysis is the same MFCC-analysis as in ASR-ELIS.

The phonological feature extraction is different. The vectors X_{t-1}, X_t and X_{t+1} are converted into 14 distinct phonological features describing voicing, place of articulation, turbulence, nasality, etc. We now extract only phonological features that are revealed by local information since we are not interested in modeling the phonetic structure of the utterance here. This means that a modulation feature like "trill" and a transient feature like "burst" are currently not considered.

The phonological feature extractor is composed of 25 MLPs which have been trained on the read speech part of CoGeN. To prepare the training data, we first create a table containing the canonical values of the 14 phonological features of each phone. Nine phonological features, like "nasal" for instance, can be on/present, off/absent or irrelevant and are modeled by a tandem of two MLPs: one that distinguishes between relevant (=1) and irrelevant (= 0) and another that distinguishes between on (= 1) and off (= -1). The latter MLP also takes the output of the first MLP into account. Two other features, namely the vowel properties "front-back" and "high-low", are also modeled by two MLPs. Since pathological speakers mainly experience problems to realize the vowels at the extremes of the vowel trapezium rather than the central vowels, the first MLP distinguishes between "central" or "not a vowel" (= 0) and "non-central vowel" (= 1) while the second MLP distinguishes between "front" or "high" (= 1) and "back" or "low" (= -1) respectively. Finally, three binary features "voicing", "silence" and "turbulence" are modeled by a single MLP.

Given that there are 11 ternary and 3 binary features, the output Y_t consists of 25 continuously valued components $Y_{ti}, i = 1, ..., 25$, each representing the degree of confidence for the presence/absence and the relevance/irrelevance of one phonological feature at frame t.

We hypothesize that the fluctuations over time in a phonological feature, called a feature pattern, can reveal an articulatory deficiency of the speaker, in spite of the fact that the phonetic nature of the frames is unknown (that knowledge would have to come from an ASR). If this hypothesis holds, it should be possible to derive relevant features from a statistical analysis of the feature patterns Y_t. Obviously, this may not be true anymore if the utterance is too short to have a phonetic content that is sufficiently representative of speech in general.

Although an analysis of subvectors of Y_t could be interesting as well, we restricted ourselves to an analysis of individual features Y_{ti}. If a component of Y_t either describes a binary feature or the relevance of a ternary feature, the statistical analysis runs over all frames. If it is the presence/absence of a ternary feature, the analysis runs over the frames with a positive relevance. We derive both frame-level and segment-level statistics. To that end we define relevant (irrelevant) segments as intervals of at least 4 consecutive frames where a ternary feature is relevant (irrelevant). Similarly, we define positive (negative) segments as intervals of at least 4 frames where a relevant feature is present (absent). For every component of Y_t,

the following features are derived:

1. mean value,
2. standard deviation,
3. percentage of relevant/positive frames,
4. percentage of relevant/positive segments,
5. mean over all relevant/positive frames,
6. mean over all irrelevant/negative frames,
7. mean duration of a relevant/positive segment,
8. mean duration of an irrelevant/negative segment,
9. mean of the peaks (maxima) in the relevant/positive segment,
10. mean of the valleys (minima) in the irrelevant/negative segment,
11. mean time needed to reach the maximum within a relevant/positive segment,
12. the mean time needed to reach the minimum within an irrelevant/negative segment.

A block diagram of the phonological feature extraction can be found in Figure 8.1.

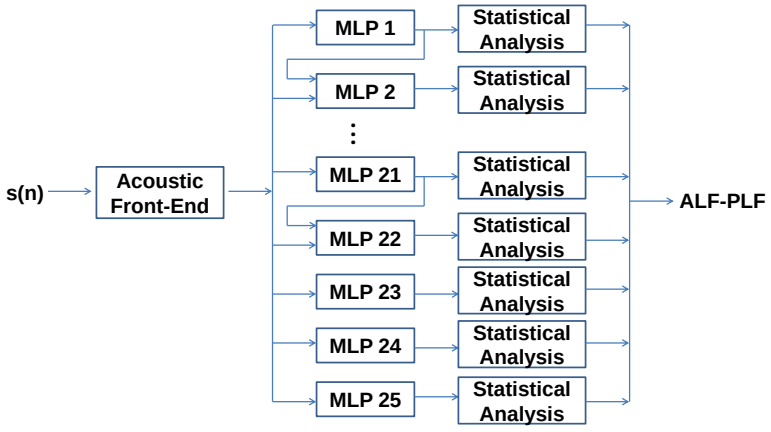

Figure 8.1: Block diagram of the phonological feature extraction process.

Most features aim to reveal whether the speaker has difficulties in realizing clear presence/absence/irrelevance distinctions, but others are more looking for problems related to the switch between presence and absence. In total, a speaker is characterized by 25 x 12 = 300 features. Obviously, one can expect high correlations between some of these features.

8.3.2 Alignment-free phonetic features

The ALF-PLF are expected to be powerful if the intelligibility reduction due to a certain speech disorder can be attributed to problems with the realization of individual phonological classes. Nevertheless, it may well be that this degradation mainly follows from problems that only arise when a certain combination of phonological classes must be realized, e.g. the realization of "voicing" and "fricative" in phone /z/. In that case, intelligibility prediction could benefit more from features that take these interactions between phonological classes into account. An obvious way to accommodate this is to consider the phons employed for deriving the PLF as phonological class combinations and to examine how much evidence for these combinations can be found in the speech utterance. .

In order to derive the envisioned phonetic features, we start from the phonological description that is created by means of the PFD that was included in ASR-ELIS as described in Section 7.1.2. The reason for this is that while the ALF-PLFs only need a small context window, more context is needed to derive phonetic features. Using Equation (7.3), we assign each frame to the phone $F_k, k = 1, ..., N_F$ yielding the largest probability

$$P(F_k|X_{t-5,..,t+5}) = \left[\prod_{\substack{i=1 \\ A_{ci}(F_k)=1}}^{24} Y_{ti} \right]^{\frac{1}{N_p(F_k)}} \tag{8.1}$$

with $N_p(F_k)$ being the number of positive phonological feature characteristics of F_k.

We then consider all frames that were assigned to a particular F_k and we derive four features of that F_k: the mean, the standard deviation, the mean of the valleys and the mean of the peaks of the $P(F_k|X_{t-5,..,t+5})$. These four features investigate how well the speaker's utterances of phones which are perceived as F_k match the acoustic models for F_k. What we also want to know is how well the phone F_u the speaker really wanted to utter is recognized. To this end, we define $P(F_k, F_u, F_r|U, R)$ as the probability that F_k came out as the winner (largest posterior probability according to Equation (8.1)) while the speaker actually tried to utter F_u although according to the text it had to be F_r. Based on this definition, we can express the probability that F_k is the winner, given the utterance and the text, as

$$P(F_k|U,R) = \sum_{u,r=1}^{N_F} P(F_k, F_u, F_r|U, R) \qquad k = 1, .., N_F \tag{8.2}$$

$$= \sum_{u,r=1}^{N_F} P(F_r|R)P(F_u|F_r)P(F_k|F_u) \quad k = 1, .., N_F \tag{8.3}$$

The meanings of the probabilities in the right hand side are the following:

- $P(F_k|F_u)$ is the probability that F_k is the winner when the speaker tries to utter F_u. Obviously it depends on the quality of the phonological analyzer, but more importantly, on the difficulties the speaker experiences to pronounce F_u.
- $P(F_u|F_r)$ is the probability that the speaker tries to pronounce F_u when according to the canonical transcription of the text it should have been F_r. It is a measure of how many times the speaker is making a reading error.
- $P(F_r|R)$ is the probability that F_r appears in the canonical transcription of the text. This is strictly a property of the text, but if the text is long enough it will be more like a property of the language.

If we assume that the speaker only makes very few reading errors, Equation (8.3) can be simplified to

$$P(F_k|U,R) \simeq \sum_{r=1}^{N_F} P(F_r|R)P(F_k|F_u = F_r) \tag{8.4}$$

Apparently, what we can measure, namely $P(F_k|U,R)$, is a weighted sum of the $P(F_k|F_u = F_r)$ we would really like to measure. Moreover, the weights in that sum depend on the properties of the text. In order to suppress that dependency, we propose to compute the scaled posterior

$$\frac{P(F_k|U,R)}{P(F_k|R)} \simeq \sum_{r=1}^{N_F} \frac{P(F_r|R)}{P(F_k|R)} P(F_k|F_u = F_r) \tag{8.5}$$

If the acoustic models achieve a sufficiently good performance, one can expect that only those combinations (F_k, F_r) corresponding to confusable pairs will contribute to that sum. Furthermore, since the MLPs are trained in a discriminative way, F_r will only be confused frequently with F_k if the prior probability of F_k is at least as high as that of F_r. This actually means that one can argue that the dominant terms in the sum will have a ratio $P(F_r|R)/P(F_k|R)$ that is close to 1, and thus that the text-dependency will be low. Consequently, we propose to add the scaled posteriors of the different phones as a new set of alignment-free features. These posteriors $P(F_k|U,R)/P(F_k|R)$ can be calculated in two ways. One way is to consider per F_k the percentage of frames for which F_k was the winner as an estimation of $P(F_k|U)$, and divide that percentage by $P(F_k|R)$. A second way is to use the mean $P(F_k|Y_t)$ over the whole utterance as an approximation for $P(F_k|U)$ and divide this number by $P(F_k|R)$. These two approximations constitute the fifth and sixth alignment-free feature of F_k. This way, we finally obtain $6N_F = 6 * 55 = 330$ new alignment-free phonetic features. Since most of the

phones represent phonemes, we denoted these features as ALF-PMF in order to maintain the duality between features associated with phonological classes and features associated with phone(me)s.

8.3.3 Results on COPAS

Returning to the RSI prediction of 121_TM, we built IPMs based on the two new feature sets ALF-PLF and ALF-PMF. We also evaluated some straightforward combinations, such as ALF-PLF + ALF-PMF (all alignment-free feature sets together), PLF + ALF-PLF, the phonological feature sets together, and PMF + ALF-PMF, the phone(m|t)ic feature sets together. Table 8.2 presents the results of these experiments.

feature set	RMSE	PCC	learner
PLF	8.80	0.75	ELR
PMF	9.20	0.72	ELR
PLF + PMF	**8.57**	**0.77**	**ELR-late**
ALF-PLF	9.89	0.64	ELR
ALF-PMF	9.23	0.71	ELR
ALF-PLF + ALF-PMF	8.94	0.73	ELR-early
PLF + ALF-PLF	8.88	0.74	ELR-late
PMF + ALF-PMF	8.69	0.76	ELR-late

Table 8.2: RMSE and PCC between computed and perceptual phoneme intelligibility scores based on alignment-based and alignment-free methods. Results differing significantly at a level of $p < 0.05$ from the reference PLF+PMF model (in bold) are underlined.

While for the alignment-based methods PLFs lead to stronger IPMs than PMFs, this is not the case anymore for alignment-free methods. ALF-PLFs are outperformed by the ALF-PMFs, although not significantly. Results based on the latter feature set are not significantly worse than those attained with the ASR-based feature combination PLF+PMF. This makes the ALF-PMF set a potential candidate for further use. However, they don't seem to add complementary information to the ASR-based methods since PMF + ALF-PMF does not improve on the best results. Both alignment-free feature sets together result in an IPM performing almost not significantly worse than the best alignment-based combination. Since alignment-free feature sets are less complex, need less CPU time and are supposed to be less dependent the text, they offer a valid alternative for the alignment-based approach.

8.4 Comparison with PEAKS

As described in Chapter 4, PEAKS is a German tool that was designed for automatic intelligibility assessment of pathological speech. Since our research has the same objectives, it is worthwhile to compare our strategy with that of PEAKS.

As opposed to the forced alignment method in the Computerized Frenchay Dysarthria Assessment, the creators of PEAKS opted for a word recognition approach [3, 110]. The recognizer used for adult speech assessment is designed for achieving a maximal performance on readings of a particular text passage which is the same for all speakers. For intelligibility assessment, this is the "Nordwind und Sonne" passage, a text which is frequently used in speech therapy [7] in German speaking countries. It is composed of 108 words (71 disjunctive) and phonetically balanced (it contains all phonemes of the German language). The HMM-based ASR is trained on non-pathological German speech and the lexicon consists of the words appearing in the targeted text passage and their expected pronunciations. The language model is an unigram model comprising the frequencies of occurrence of the words in the text passage. Sentence per sentence, the outputs of the ASR are compared with the targeted sentence, and from these comparisons, a Word Accuracy (WA) is derived:

$$WA = 1 - \frac{D + S + I}{R}, \tag{8.6}$$

with D the number of deletions, insertions I and substituted words S versus the number of words R in the reference.

It was soon acknowledged that WA alone is not enough to achieve good intelligibility predictions. Therefore, PEAKS also computes a number of acoustic and prosodic features that characterize the speech of the speaker, like e.g. the mean and variance of F_0 of each recognized word, and some statistics about jitter, shimmer and the number and length of the voiced and unvoiced segments over the whole utterance. The way these speaker features are computed exhibits some resemblance with how we do it, in the sense that the word segmentation created by the recognizer is taken into account, but for the rest there are only few points of similarity. Some of the PEAKS features describe F_0-patterns in words and are in a sense related to the AMPEX features that we will be using for phonation quality prediction later in Chapter 9. Other PEAKS features describe the number of voiced and unvoiced segments encountered in the utterance and are somewhat related to some of the alignment-free features we proposed in this work.

To map the speaker features to intelligibility scores, PEAKS embeds an IPM which is based on SVR.

8.4.1 Published performances of PEAKS

As described in [3], PEAKS was mainly tested on (partially) laryngectomized adults and children with cleft lip and palate. We will focus on the first group here since we ourselves do not dispose of running speech data for children to compare against the German data. The IPM for (partially) laryngectomized adults is supplied with the speaker features described above. It was trained and evaluated on a group of 41 laryngectomees with TE-speech who read the "Nordwind und Sonne" passage. For every patient, 5 voice professionals provided perceptual RSI scores, expressed on a 5-point Likert scale [65]. These 5 scores per patient were averaged to one reference score, being the ground truth for training of the IPM.

The training and evaluation is obtained by means of a leave-one-out procedure which can be described as follows [3]:

- Consider all speakers except one as training speakers and the left-out speaker as the test speaker.
- Select the n features showing the highest correlation with the reference scores on the training dataset.
- Use this feature subset to train a SVR on the training set and validate it on the left out speaker to determine the prediction accuracy.
- Repeat the above steps for every speaker and keep track of the predicted speaker intelligibilities.
- Compute the PCC and the Spearman rank correlation coefficient [127] between predictions and reference scores to determine the prediction accuracy.

Starting from $n = 1$, the number of features is augmented until the prediction accuracy on the leave-one-out samples does not further improve anymore. Note that the best results of this approach will be optimistic since the test data are used to find the free parameter n.

Following the sketched approach, high correlations between the reference and the automatic scores [3], were found: while the inter-rater PCC was between 0.80 and 0.87, the PCC between the mean perceptual and the automatic scores was around 0.90.

In a separate experiment, Maier et al. [3] tried to predict scores emerging from one single rater instead of from the mean of five ratings. Using the same strategy as before, the PCC between one rater's scores and the automatic scores dropped to values between 0.74 and 0.80 (depending on the targeted rater), which is slightly under the inter-rater PCC.

8.4.2 Comparison with our former experiments

Although this is not offering a direct comparison, we can consider the PI predictions we obtained with pathology-specific models derived from running speech on the group of laryngectomees in COPAS. Taking into account that our ground truth was phoneme intelligibility and that it was delivered by a single human rater, we argue that our results provide evidence that our strategy is at least competitive with the PEAKS approach. Furthermore, it has been tested on a much more diverse population of disordered speech (the full COPAS), whereas there is no proof yet that PEAKS will also scale up to that degree of diversity.

To gather more conclusive evidence concerning the relative performances and the degree of complementarity of the two strategies, we should be able to apply both methods on the same dataset and to investigate whether a combination of methods can improve on the single model performances. This was one of our motivations for seeking collaboration with the group of Erlangen. Our collaborations with third parties are discussed in the next section.

8.5 Collaborations with third parties

As mentioned earlier, we introduced two novel alignment-free feature sets for characterizing a speaker. One of them, ALF-PLF, is the result of a phonological analysis of the speech which does not need any knowledge of what has been said. The other one, ALF-PMF, is also the result of a phonological analysis, but this analysis relies on the relations between phonemes and phonological classes and needs phoneme frequencies that are being retrieved from the text the patient was supposed to have read (the prompt). Since the alignment-free features seem capable of well predicting phoneme intelligibility in COPAS, and since the link with what has been spoken is rather weak, the method should be easily transferable to other languages. In order to verify this, we have collaborated with the Chair of Pattern Recognition of the University of Erlangen (LME). The results of that analysis will be discussed in Section 8.6, and have been published in [19].

Similarly, a collaboration with the Netherlands Cancer Institute and the Institute of Phonetics of Amsterdam University (NKI-UVAFON) was set up in order to investigate issues like text dependency, accent dependency and predictability of articulatory variables other than intelligibility. The first results of that collaboration will be discussed in Section 8.7 and have been published in [14]. The rest of these collaboration results will be discussed in Chapter 9.

8.6 Collaboration with LME

LME and ELIS are both research groups with a good background in pathological speech analysis. While the DIA online tool was developed in ELIS [17], LME developed the PEAKS-platform [3]. In both PEAKS and DIA, the speaker's utterance is analyzed in view of the prompted text.

In PEAKS, this analysis is made by an ASR with a small dictionary, limited to the words in the prompted paragraph. The basic idea is that the ASR system has increasing trouble recognizing pathologic speech with an increasing degree of pathology. Intelligibility is measured as the percentage of correctly recognized words.

In DIA, the speech is aligned with the list of prompted words, and from that alignment a set of speaker features is retrieved and subsequently transformed into an objective intelligibility score. As much as the above methods have proven to work well for the task they were designed for, they are presumed to run into problems when the pathological speaker starts to make hesitations and reading errors, as it often happens with children speakers.

Clearly, these errors should have no impact on the intelligibility, but they do introduce out of vocabulary words which may on their turn cause an alignment or a recognition system to derail. To circumvent this lexical problem, a new philosophy of deriving speaker features was conceived almost simultaneously in LME [126] and ELIS [18]. In this philosophy, no alignment (needing the prompts) nor recognizer (needing a lexicon) is employed anymore.

A first attempt was made by Bocklet et al. [126] where a speaker verification approach is adopted: a GMM is trained for every speaker, and the parameters of that GMM constitute a supervector from which to predict the speaker's intelligibility. This supervector represents the acoustical properties of the speech. This method led to high correlations between computed and perceptual intelligibility scores for a German dataset consisting of 85 partially laryngectomized speakers. As only acoustical properties of the speech are used, this approach is claimed to be language-independent.

In the research work described in the previous sections an alignment-free feature set ALF-PLF, relying on nothing but a statistical analysis of the feature patterns emerging from phonological feature detectors was introduced and showed promising results on COPAS. The phonological feature set is claimed to be independent of the used language as well. Moreover, it is presumed to relate directly to the articulatory dimensions of speech, and as such, possibly suitable for conducting a more detailed assessment of the speaker's articulation problems in a later stage (see Chapter 9).

As both ASR-free approaches capture different characteristics of the speech

signal, it makes sense to verify the claim of language-independence and to investigate whether combining them is beneficial. For this purpose, we conducted experiments on the two datasets that were formerly used to test the individual approaches as presented in [126] and [18] respectively.

8.6.1 Datasets

In this subsection we describe the two datasets we used for training and evaluation of the individual and combined models.

8.6.1.1 German Partial Laryngectomees (GPL)

The dataset used in [126] contains recordings of 85 patients who suffered from cancer in different regions of the larynx. 65 patients had already undergone partial laryngectomy and were recorded on average 2.4 months after surgery, while the remaining 20 patients were still awaiting surgery. Each person read the German version of "The Northwind and the Sun". More details about the recording conditions can be found in [126].

Five phoneticians and speech scientists rated every speaker's intelligibility according to a 5-point Likert scale [65]. The average of these five ratings is used as a reference during the automated intelligibility assessment. Pearson Correlations between scores of one rater and the average scores of the four others range from 0.76 to 0.86, with a mean of 0.81.

8.6.1.2 Flemish Pathological Speech (FPS)

This dataset is nothing else than the 121_TM set from COPAS. Here the name points to the fact that the database is Flemish, in contrast to the GPL. As already explained, perceptual phoneme intelligibility (PI) scores (derived from the DIA recordings) are available for all speakers, but there are no perceptual running speech intelligibility (RSI) scores. Here we consider the PI score as a proxy for the RSI score. Since we do not dispose of exact inter-rater agreements for this dataset, the estimations for inter-rater agreements mentioned in Chapter 7 (PCC of 0.89) were taken as a reference.

8.6.2 Feature extraction

For both derived feature sets, the same acoustic front-end was used, computing the standard MFCCs with a frame rate of 10 ms and a frame size of 25 ms. For each frame t, the first 12 MFCCs and the log energy are retained. LME also uses first and second order derivatives, to constitute a 39-dimensional feature vector X_t. To minimize the influence of the microphone, Cepstral Mean Subtraction (CMS) is applied to all data.

Based on the acoustic features, the approaches of [126] and [18] are applied to create two alignment-free speaker feature sets, representing the acoustical and phonological properties of the speech respectively.

8.6.2.1 Acoustical alignment-free features (ALF-AC)

The first system, described in [126], is based on the assumption that the acoustics of pathologic speakers differ from those of non-pathological speakers. The degree of pathology is measured as the distance between the pathologic speaker model and a reference speaker model. The speaker model is a Gaussian Mixture Model (GMM) representing all available MFCC vectors $\mathbf{X}$ of the speaker. The reference model is a speaker-independent GMM that is trained on speech of healthy speakers. This model is usually referred to as the *Universal Background Model* (UBM).

The UBM is trained in an unsupervised iterative manner by the *Expectation-Maximization* (EM) algorithm [88] (5 iteration steps). It computes likelihoods by means of Equation (5.27) and its free trainable parameters are the weights, mean vectors and covariance matrices of the different mixtures. The number of Gaussian mixtures is set to 128.

The speaker model of the diagnosed speaker is derived by adapting the parameters of the UBM to the data of the speaker. Since only a limited amount of data per speaker is available, only the mean vectors are adapted. This is accomplished by means of *Maximum A Posteriori* (MAP) adaptation [88]. The adapted means constitute a so-called GMM-based supervector (see Figure 8.2). This vector is expected to represent well the acoustic space of the speaker. It is referred to as ALF-AC and it is composed of 39 x 128 = 4992 individual features.

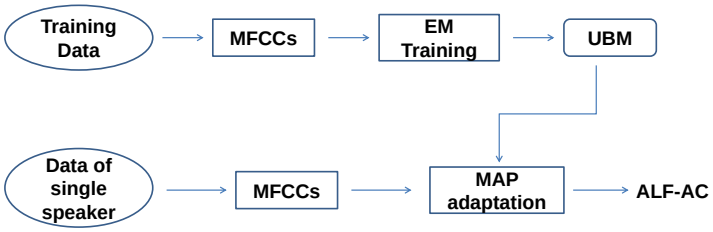

Figure 8.2: Composition of the GMM-based supervector by concatenation of the mean vectors. After [19].

8.6.2.2 Phonological alignment-free features (ALF-PLF)

While the ALF-ACs describe the speaker in the acoustic space, the ALF-PLF describe him/her in the phonological feature space. The ALF-PLF feature set, de-

scribed in 8.3.1, consists of 300 features.

8.6.3 Experimental setup

For experiments on FPS, the UBM for the ALF-AC methodology is trained on all Flemish utterances of the TM paragraph. MAP-adaptation towards the tested speaker leads to the speaker feature vector. The ALF-PLF methodology embeds a phonological detector that was trained on "normal" Flemish speech as described in Section 8.3.1. The speaker feature vector is extracted using the usual statistical analysis.

For experiments on GPL, the UBM for the ALF-AC methodology is trained on "normal" German utterances of the same paragraph. MAP-adaptation towards the tested speaker again leads to the speaker feature vector. The ALF-PLF methodology embeds again the same phonological detector that was trained on "normal" Flemish speech as described in Section 8.3.1. No adaptations were made towards German speech. The speaker feature vector is extracted using the usual statistical analysis.

Starting from the these alignment-free speaker feature sets, four different IPMs were created per dataset. In an analogy with Figure 7.2, two of them (IPM 1 and IPM 2) consider only one feature set and act as the baseline models. The two others (IPM 3 and IPM 4) employ a combination of the two feature sets by using an early and late fusion strategy respectively.

8.6.3.1 Training and validation procedure

For the training and validation of the IPMs we adopted a leave-one-out cross validation scheme. We tried two statistical learners for every IPM: one based on ensemble linear regression (ELR) with feature selection and one based on Support Vector Regression.

Although experiments in Chapter 7 showed that SVR with a non-linear kernel (RBF) usually outperforms SVR with a linear kernel, we did keep the linear kernel as an option here because the ALF-AC feature space is a very high-dimensional space in which there may be no need to any non-linear transformation to another hidden space anymore.

The SVR experiments were conducted in Weka [128] and the learning parameters were set to the default values.

8.6.4 Results and discussion

In this subsection we present the results for the four IPMs in combination with SVR and ELR as the training algorithm. As usual, we computed the Pearson Correlation Coefficient (PCC) and the Root Mean Squared Error (RMSE) between the

computed and the target outputs as evaluation measures. The RMSE is expressed in percent of the full scale: 5 for GPL and 100 for FPS.

		SVR			ELR	
data	feature set	kernel	PCC	RMSE	PCC	RMSE
FPS	ALF-AC	LIN	0.70	9.4	0.44	11.6
	ALF-PLF	RBF	**0.68**	**9.6**	0.64	9.9
	early fusion	LIN	0.71	9.2	0.65	9.8
	late fusion	-	0.74	8.7	0.64	10.2
GPL	ALF-AC	LIN	**0.81**	**11.0**	0.72	13.0
	ALF-PLF	RBF	0.81	11.0	0.69	12.8
	early fusion	LIN	0.81	11.0	0.73	12.8
	late fusion	-	0.84	10.4	0.73	12.6

Table 8.3: PCCs and RMSEs (see text) for the two datasets. In case of SVR, linear kernels are denoted by LIN, Gaussian kernels by RBF. Per dataset, the result for the best single feature set is put in bold and acts as a reference. Underlined figures indicate performances differing significantly from the reference with $p < 0.05$.

A first major finding is that SVR outperforms ELR as a learning method in all situations. We come back to this later. Looking at the SVR models, it appears that both feature sets, ALF-AC and ALF-PLF, perform equally well on both datasets. This is proof of the fact that these two feature sets can be used in a language independent scenario, as claimed but not verified in the original papers where they were introduced. Another result is that early fusion is not capable of exploiting the complementarity of the two feature sets, whereas late fusion can, even though the improvement on GPL is only significant at a confidence level of 0.08. Optimizing the parameters of the SVR training instead of using the Weka default values and adopting a more efficient late fusion technique might further improve the results and lead to significant differences with a lower p-value on both datasets. That early fusion is not capable of causing any improvement may well be a consequence of the very unequal sizes of the combined sets.

Note that all models perform better on the GPL than on the FPS dataset when comparing to the inter-rater agreements: all GPL models achieve at least the mean inter-rater PCC of 0.81, while none of the FPS models achieves the inter-rater PCC of around 0.89. One hypothesis is that the GPL dataset only comprises laryngectomees and that the dominant cause of the diminished intelligibility of this type of speakers resides in the diminished amount of voicing that is produced. This type of deviation is obviously easier to model than a more complex articulatory deficiency involving e.g. a combination of problems related to both the manner and the place of articulation. Such complex deficiencies are bound to occur frequently in the FPS-dataset. Another factor might be the mismatch between the reference scores in the FPS dataset and the evaluated utterances. The reference scores were

PI-scores which were measured on another type of utterance (DIA word test).

The fact that the ALF-AC features perform very badly on the FPS dataset when used in combination with ELR can most likely be explained by noting that the ALF-AC feature set consists of many strongly correlated components, and by acknowledging that the simple strategy of adding one feature at the time (used in ELR) is not a viable strategy in that case. To give an example, if the mean vector of a mixture component in the speaker model differs from the corresponding mean vector of the UBM, it is probably important to measure in which direction the mean vector has moved. This direction information is encoded in a linear combination of mean vector components and is not necessarily well reflected in any of the individual components of that vector. Consequently, the feature addition method may fail to add any of these components to the subspace in which the regression will take place. In SVR, the features are always examined together. That the phenomenon is so much more apparent in the FPS dataset than it is in the GPL dataset is probably a consequence of the larger complexity of the envisaged modeling task in the FPS dataset.

8.6.5 Conclusions

Two ASR-free methods that were formerly shown to predict speech intelligibility rather well on one dataset were now compared in a multilingual setting. Both methods were evaluated on two datasets comprising Dutch and German speech respectively. The fact that both methods achieve very similar results on each of the two datasets proves that they are equally effective. The fact that each method, when used in a cross-lingual setting (=tested on data in another language than the data that were used in the original paper and, in case of the ALF-PLFs, tested on data in another language than the data that were used for the training of the acoustical models), competes with the other method in a native setting indicates that the methods can indeed be considered as language-independent as claimed but not verified in the original papers.

Another important result is that combining the two feature sets in one system demonstrated to be beneficial, provided late fusion is employed as the fusion technique. Since late fusion is achieved by just averaging the outputs of two intelligibility production models, there is still room for improvement. Future work can be directed towards the training of yet another regression model to better map the two outputs onto the desired intelligibility score.

Now that both methods have been proven to work in a Flemish-German context, we can start to explore more datasets covering more languages, in the hope that the combined method will prove to be applicable in general.

8.7 Collaboration with NKI-UVAFON

As mentioned earlier, another collaboration was set up with NKI-UVAFON, resulting in new challenges [14]. In fact, previous research in this dissertation has shown high correlations between automatically generated scores and perceptual ratings, but these correlations have always been measured on a large group of pathological speakers where the aim was to compare one speaker against another. However, in a clinical setting there is also a high need for tools that are able to monitor progress made by an individual patient. Since NKI-UVAFON had developed a corpus containing multiple recordings of the same person at 2 or 3 distinct times, one of the first aims of the collaboration was to investigate whether the developed methods are sufficiently accurate to predict the presence/absence of progress in the course of a therapy.

Since the mentioned corpus also contains recordings of two text fragments of each speaker, assessing the text-independence of our approach was another challenge.

8.7.1 The NKI-CCRT corpus

All speech material in the NKI-CCRT corpus comprises speech from patients with advanced head and neck cancer that were treated with chemoradiotherapy (CCRT, [129]). The perceptual evaluations are part of a larger longitudinal study investigating the automatic evaluation of speech intelligibility and voice quality for speakers treated for advanced head and neck cancer. Here we just provide a synopsis of the information regarding participating speakers and perceptual evaluations that have been performed on the data. We refer the reader to [129] and [130] for more detailed information.

8.7.1.1 Speakers

The corpus contains recordings and perceptual evaluations of 55 speakers: 54 of them were recorded before CCRT (T0), 48 speakers were recorded again ten weeks after CCRT (T1) and 39 speakers were recorded twelve months after CCRT (T3)[1]. As detailed in [129], the average age at pre-treatment was 57 years (range 32-79). As summarized in Table 8.4, approximately one-third of the speakers had tumors located in the laryngeal cavity (laryngo/hypopharynx). The remaining speakers had tumors located above the laryngeal cavity (oral cavity, oropharynx, nasopharynx). Based on perceptual categorization by a Dutch phonetician, 8 speakers were categorized as non-native whereas the other 47 were categorized as native. Note

[1]There were also recordings for some of the patients at time T2, situated between T1 and T3, but due to time constraints, these recordings were not perceptually rated, and therefore not used in this study. However, to maintain numerical consistency with the publication of [129], we use the term T3 for the last recording moment.

that all speakers lived in the Netherlands, and as such, their Dutch is an "accent" of Flemish.

		Evaluation moment		
	Total (%)	T0 (%)	T1 (%)	T3 (%)
Characteristic	n=55	n=54	n=48	n=39
Tumor location				
Non-Laryngeal[a]	36 (65)	36 (67)	31 (65)	23 (59)
Laryngeal[b]	19 (35)	18 (33)	17 (35)	16 (41)
Sex				
Male	45 (82)	44 (81)	39 (81)	30 (80)
Female	10 (18)	10 (19)	9 (19)	9 (23)
Lang. background				
Dutch 1st	47 (85)	46 (85)	40 (83)	32 (82)
Dutch 2nd	8 (15)	8 (15)	8 (17)	7 (18)

Table 8.4: Speaker characteristics: tumor location, sex and probable language background. [a]Tumor located in oral cavity, oropharynx, nasopharynx. [b]Tumor located in laryngopharynx or hypopharynx.

As most speakers were recorded before CCRT and at two moments after CCRT, this dataset makes it possible to monitor short-term and long-term changes in a patient's intelligibility. Preliminary results presented by [130] show however that not all speakers actually exhibit significant changes.

8.7.1.2 Stimuli

Two fragments of a 189-word passage from a Dutch fairy tale were selected as fragments A and B. Fragment A contains 70 words (tokens) while fragment B contains 68 words long. Fragment A contains 49 unique words (types) and fragment B contains 50 unique words (see Table 8.5). The two fragments have only 22 types in common, which makes them clearly lexically different. Each speaker

	Text Diversity			Sentence Length
Fragment	Tokens	Types	TTR(%)	mean (sd, range)
A	70	49	70.0	11.7 (6.3,4-21)
B	68	50	73.5	17.0 (5.8,12-23)
$A\&B$	138	77	55.8	13.8 (6.4,4-23)

Table 8.5: Characteristics of the two text fragments: number of tokens, number of types and type-per-token ratio (in %). Average sentence length is denoted in number of words. Data are rounded to one decimal place.

read at least one of the fragments, but most of them read both: the corpus contains 140 recordings of fragment A and 141 of fragment B. Average durations of the recordings were 26.9 seconds for fragment A and 26.4 seconds for fragment B.

From the phoneme frequencies in fragments A and B (see [130]) it follows that the two fragments have an almost identical phonemic balance.

8.7.1.3 Perceptual analysis

Thirteen recently graduated or about to graduate speech pathologists (all female, native Dutch speakers, average age of 23.7 years) evaluated the speech recordings in an on-line, self-paced experiment. The recordings were presented in a randomized order and listeners could replay a recording as many times as they wished. Each recording contained the reading of a complete fragment by one speaker. The listeners used their own anchors and received no feedback on performance. All listeners completed an on-line familiarization module before evaluating the stimuli for the dataset. The retest recordings (repetitions of formerly rated recording) and items for practicing are not included in the dataset.

Intelligibility was evaluated on a 7-point scale with labels provided for the scale ends ('poor' for 1 and 'good' for 7). Preliminary results presented in [130] indicate that although some listener's test-retest reliability was low, the Interclass Correlation Coefficient [67] assessing the between-rater reliability was 0.95 (based on a sample of 37 items). This high value indicates that the mean intelligibility scores (averages over listeners) are reliable. The percentage exact agreement for the rater's test-retest recordings ranged from 20 to 80 percent. The percent close agreement ($\pm$ 1 difference on the scale) ranged from 60 to 100 percent. In terms of Pearson Correlation Coefficient (PCC), the correlation between one individual rater and the mean of the 13 raters varies between 0.72 and 0.92, with a mean of 0.84. Figure 8.3 depicts the histogram of the mean perceptual ratings for all recordings.

8.7.2 Objective intelligibility assessment

The derivation of an objective intelligibility score is the same as always and consists of the same front-end analysis (derivation of acoustic parameter vectors), a speaker feature extraction and an intelligibility prediction.

8.7.2.1 Speaker feature extraction

Since one of our aims is to investigate the accent dependency of our method, we have developed two sets of acoustic models: models trained on Flemish speech and models trained on Dutch speech. The Flemish models are trained as discussed in Chapter 7: the Flemish models for ASR-ELIS are trained on the read speech

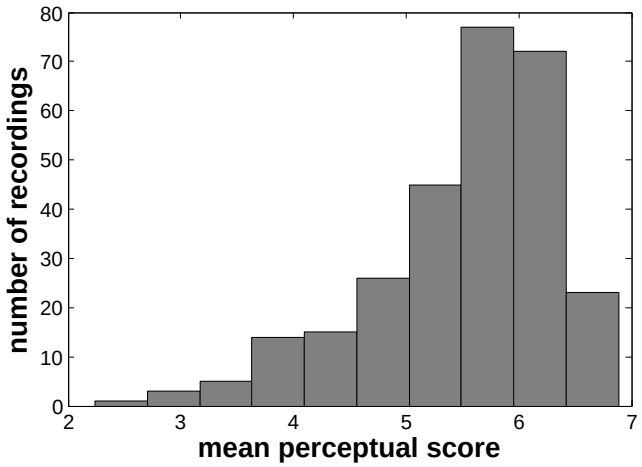

Figure 8.3: Histogram of the mean perceptual intelligibility scores in the NKI-CCRT corpus.

part of CoGeN, those for ASR-ESAT on the Flemish read speech part of CGN. The Dutch models are all trained on the Dutch read speech part of CGN.

Flemish and Dutch are known to differ e.g. in the voicing of fricatives and the degree of diphtonguation of long vowels [131, 132]. These differences must be taken into account when assigning canonical values A_{ci} to the phone states (e.g. the /g/ is pronounced with voicing in Belgium and can be either voiced or unvoiced in the Netherlands). This is relevant for the PLF and the ALF-PMF feature sets.

In summary, we developed a Flemish version and a Dutch version of PLF, PMF, ALF-PLF and ALF-PMF. The Flemish version of feature set $\mathcal{F}$ will be noted as FL-$\mathcal{F}$, the Dutch version as DU-$\mathcal{F}$.

8.7.2.2 Intelligibility Prediction Model

Once all speaker features have been computed, they need to be converted to an intelligibility score using a regression model, the Intelligibility Prediction Model (IPM). We chose to use ELR as the only statistical learner this time since this offers the possibility to identify the important dimensions for describing intelligibility.

8.7.3 Experimental evaluation

The main objectives of the experimental evaluation are to assess the accuracy of the IPMs derived from the different speaker feature sets and their robustness against changes in the read text and the spoken accent (Dutch or Flemish). In order to reach these objectives we have derived IPMs from different text fragments (frag-

ments A and B). To investigate accent dependency, we tested feature sets derived by means of acoustic models trained on Flemish and Dutch normal speech respectively. In case of acoustic models working in the phonological feature space, the phonological description of the phone states was matched to the speech material that was processed: Flemish descriptions for processing Flemish speech and Dutch descriptions for processing Dutch speech and testing on the NKI-CCRT corpus.

Before describing our experimental results in more detail, we first take a closer look at the evaluation strategy we have adopted. All IPMs will be trained and evaluated using a 5-fold cross validation (CV) strategy. As most speakers were recorded two or three times (at T0, T1 and/or T3) and since two fragments (A and/or B) were recorded in most cases, 281 samples were available in total. These samples were divided into five folds such that all recordings of one speaker are always placed in one fold. Performance is again expressed in terms of the RMSE and PCC. For all patients, the average of the thirteen intelligibility ratings is taken as the reference score.

8.7.3.1 Individual speaker feature sets

In a first experiment we tested the Flemish and Dutch variants of the four feature sets we proposed and we used these feature sets in combination with IPMs that were trained and tested on the same fragment. In view of later experiments we introduce the notation $A \rightarrow B$ for instance to express that the IPM is trained on fragment A and tested on fragment B. The results for $A \rightarrow A$ and $B \rightarrow B$ can be found in Table 8.6.

	fragment A				fragment B			
	FL		DU		FL		DU	
features	RMS	PCC	RMS	PCC	RMS	PCC	RMS	PCC
PMF	0.82	0.60	**0.65**	**0.77**	0.68	0.73	**0.60**	**0.77**
PLF	0.83	0.58	0.79	0.60	0.75	0.63	0.68	0.72
ALF-PLF	0.77	0.63	0.77	0.62	0.74	0.66	0.73	0.66
ALF-PMF	0.68	0.73	0.68	0.73	0.70	0.70	0.70	0.70

Table 8.6: Performances of IPMs departing from Flemish (FL) and Dutch (DU) feature sets and being trained and tested on text fragments A and B respectively. Per fragment, results differing significantly at a level of $p < 0.05$ from the best result (indicated in bold) are underlined. RMS denotes RMSE.

The main conclusion is that the phone(m|t)ic features (PMF and ALF-PMF) outperform the corresponding phonological features in most cases. There is only one exception to this rule, namely the Flemish PLF performing worse than ALF-

PLF on fragment A. The difference between PLF and PMF can be partly explained by differences in the systems supplying the text-to-speech alignments that are needed for constructing the speaker features: the state-of-the-art ESAT-ASR usually leads to a better alignment than the much less complex ELIS-ASR. The difference between ALF-PMF and ALF-PLF on the other hand cannot be explained in terms of the alignment (there is none) nor in terms of the phonological analyzers that were used (they were actually very similar). The data seem to support the hypothesis that intelligibility reductions are more correlated with co-occurrences of phonological classes, as they materialize in specific phonetic units, than with individual phonological classes.

The second conclusion we can draw is that the alignment-free features have a more consistent performance across different configurations than the alignment-based features. On the other hand, the alignment-based features do usually lead to the highest performance (again with the exception of Flemish PLF on fragment A). The latter is due to the fact that the speakers recorded in the NKI-CCRT corpus were mostly native adults who did not make many reading errors which could have derailed the alignment.

8.7.3.2 Robustness against the speaker accent

A very striking result with respect to the impact of the speaker accent is that the alignment-based methods are sensitive to a change of accent whereas the alignment-free methods are not. As expected, the alignment based models clearly perform better when the acoustic models are matched to the accent of the speaker.

A possible explanation for this is that the alignment is better when it is achieved with matched models, but that the statistical analysis (conducted to retrieve alignment-free parameters) is not much affected by the fact that the models are sub-optimal.

8.7.3.3 Robustness against changes in the text

In order to investigate this aspect we have conducted an additional experiment in which we tested the matched alignment-based feature sets (DU-PMF and DU-PLF) and the matched alignment-free feature sets (DU-ALF-PLF and DU-ALF-PMF) in combination with matched and unmatched IPMs. The IPM is called unmatched if it was trained on recordings of text material that was different from the text read during model evaluation. More specifically, we trained IPMs on recordings of fragment A and evaluated them on recordings of the same fragment A and on recordings of another fragment B. We did the same with IPMs trained on recordings of fragment B. The results of this experiment can be found in Table 8.7. The data clearly demonstrate that all feature sets show the same performance on a particular fragment, irrespective of whether the IPM was trained on the same or on another text. However, the used test set does play a role. The differences between

	$A \rightarrow A$		$B \rightarrow A$		$B \rightarrow B$		$A \rightarrow B$	
feature set	RMS	PCC	RMS	PCC	RMS	PCC	RMS	PCC
DU-PMF	0.65	0.77	0.68	0.76	0.60	0.77	0.60	0.76
DU-PLF	0.79	0.60	0.80	0.61	0.68	0.72	0.66	0.72
DU-ALF-PLF	0.77	0.62	0.77	0.62	0.73	0.66	0.73	0.65
DU-ALF-PMF	0.68	0.73	0.70	0.72	0.70	0.70	0.71	0.70

Table 8.7: IPMs developed on fragment X (A or B) and tested on fragment Y (A or B) as indicated by the notation $X \rightarrow Y$.

the figures obtained by testing on A and B are much larger for the alignment-based than for the alignment-free feature sets. This proves that the latter feature sets are more robust to changes in the text during evaluation. We argue that this stems from the fact that the quality of the alignment depends to some extent on the phonetic content of the text (it is known that some sound sequences are much more difficult to segment than others). The mismatch between the training and the evaluation text does not seem to be a problem.

8.7.3.4 A combination of speaker features

From the former experiments it follows that none of the speaker feature sets leads to a correlation between the objective and the perceptual scores that can compete with the mean correlation of 0.84 observed between individual raters and the mean of these raters. In this respect, we have investigated whether the combination of different feature sets may bridge this gap. We have tested combinations of:

- DU-PLF and DU-PMF (the Dutch version of the best combination for RSI as found in Section 8.2),
- the two alignment-free feature sets DU-ALF-PLF and DU-ALF-PMF,
- the two phonological feature sets DU-PLF and DU-ALF-PLF, and
- the two phone(mic) feature sets DU-PMF and DU-ALF-PMF.

The results obtained with these combinations are listed in Table 8.8.

Clearly all feature set combinations perform better than the individual feature sets they are composed of, but in most cases the improvement is not significant. The combination of DU-PMF and DU-ALF-PMF on the other hand leads to significantly better results than the separate feature sets and it yields a PCC is as high as the human inter-rater correlation. Figure 8.4 shows a convincing scatter plot of the objective versus the perceptual scores emerging from the IPM designed for this combination.

feature combination	RMSE	PCC
DU-PLF + DU-PMF	0.61	0.80
DU-ALF-PLF + DU-ALF-PMF	0.68	0.73
DU-PLF + DU-ALF-PLF	0.64	0.74
DU-PMF + DU-ALF-PMF	**0.52**	**0.85**

Table 8.8: Predictive power of IPMs built on different combinations of two feature sets. Listed are RMSE and PCC between the computed results and the means of the 13 perceptual raters. Underlined results differ significantly $(p < 0.05)$ *from the best result, denoted in bold.*

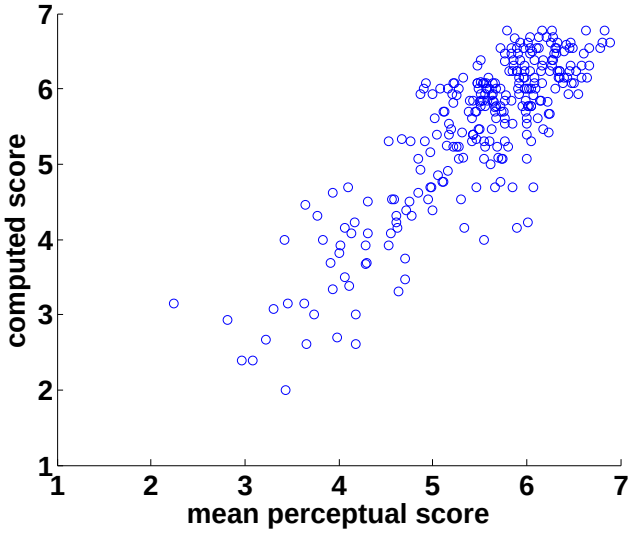

Figure 8.4: Correlation between perceptual and computed scores.

For the best combination we have also investigated in more detail how many features and which features were selected. As we adopted a five-fold cross validation strategy, 5 models were created and each of these models was on its turn obtained as a combination of 10 small models each selecting 7 - 8 features. On average, the combined model incorporated 25 features (range 21-29). Per fold, statistics were calculated on how many times a feature was selected in one of the ten small models. Features selected 5 times or more are the PMF /r/,/A/,/@/,/i/ and the ALF-PMFs /A_min/ and /N_max/, where /A/ is the vowel in the Dutch word "man", /@/ stands for the schwa in "de" and /i/ is the long vowel of "tien". Furthermore, /A_min/ is the mean minimal posterior probability for /A/ and /N_max/ is the mean maximum posterior probability for /N/, which is the final nasal sound of the word "koning". Apparently, four out of these six features are vowel-related. If we take a closer look at them, we observe that these vowels define the diagonal

of the vowel trapezium in the (place,height) plane: /i/ determines the upper-left corner (as it is *front* and *high*) while /A/ determines the lower-right corner (as it is *back* and *low*) and /@/ represents the center of this diagonal. Consequently, the vowel features can represent the amount of variation from the neutral (central) position the speaker can achieve in two directions. Together they represent the size of the speaker's vowel trapezium as a potential factor affecting his intelligibility.

Although few articles describe the speech of people treated with chemo-radiation therapy, it is known that even chemo-radiation therapy affects the organic structures and tissues around the tumor location [129]. Swallowing problems are common, and tongue and palate tissues are affected at least for part of the speakers. Persons with reduced tongue motility are known to show a strong correlation between intelligibility and vowel trapezium size [115, 133].

The fact that tongue motility can be affected in this patient group also explains the selection of features /r/ and /N_max/ as realizations of the uvular /r/ and /N/ need good functioning of the back of the tongue. Secondly, [129] shows that nasality is significantly worsened by CRRT treatment. Nasality is thus an issue in our dataset, and it is not so surprising then to notice that a nasal related feature such as /N_max/ is selected.

8.7.3.5 Patient monitoring

Now that we have established an IPM that can mimic evaluations made by a group of listeners for the comparison of one speaker against another, the next challenge is to prove that this model can also track changes in an individual patient's intelligibility over time.

First of all we have investigated whether such trends are exposed by the data. To that end we have determined the differences between the ratings of the same fragment read by the same speaker at times T0 and T1, T1 and T3 and T0 and T3. A former analysis of these data [130] demonstrated that not all speakers show a clear trend (neither progress nor deterioration) over time. For each speaker we computed the score differences at times T0 and T1, T1 and T3 and T0 and T3 and then we computed the PCC between the difference emerging from the scores of one individual rater and that emerging from the mean scores over all raters. As revealed by Table 8.9, the overall PCCs are rather low. Since we did not expect our IPM to outperform human raters, we selected those speakers for which the human raters seemed to agree on the presence and direction of the trend. The correlations between one rater and the mean of the 13 ratings for these speakers are listed in Table 8.9, together with the number of recordings for which this is the case. Based on the inter-rater agreements listed in Table 8.9, we can conclude that one can only measure a clear trend from T1 to T3 for 8 speakers. As this is considered insufficient to measure reliable correlations, we will only consider T1-T0 and T3-T0. The results of this analysis are listed in Table 8.10. In the case

	all trends			only clear trends		
times	mean	range	number	mean	range	number
T1-T0	0.56	0.45 - 0.70	93	0.70	0.40 - 0.84	26
T3-T1	0.44	0.17 - 0.62	74	0.75	0.43 - 0.96	8
T3-T0	0.62	0.45 - 0.75	78	0.78	0.60 - 0.89	28

Table 8.9: Inter-rater agreements (PCC) on speaker trends measured on all trend data and on the data exhibiting a clear trend.

	all trends			only clear trends		
times	**IPM**	mean	range	**IPM**	mean	range
T1-T0	**0.41**	0.56	0.45 - 0.70	**0.51**	0.70	0.40 - 0.84
T3-T0	**0.62**	0.62	0.45 - 0.75	**0.82**	0.78	0.60 - 0.89

Table 8.10: Correlations on speaker trend level. Results from the IPM are marked in bold.

of T3-T0, the mean human-machine-correlation is as good as the mean correlation between one rater and the mean rating, and even better for the clear trends. For T1-T0, the mean correlation between humans and the mean perceptual score is higher, but nevertheless, the human-machine correlation is in the range of human correlations, at least for the cases with a clear trend. We can therefore conclude that the IPM we developed seems able to follow the progress of an individual speaker as (un)reliably as human raters can.

Figure 8.5 shows the means and standard deviations of the T3-T0 differences in the human ratings for the 26 cases that were categorized as exhibiting a trend. Also on the Figure one finds the predicted trends. There is a lot of uncertainty on the human ratings but for 10 out of 13 of the subjects exhibiting a negative trend, the model also predicts a negative trend. The positive trends are less pronounced and, likewise, not so well predicted. Needless to say that the plot for the T1-T0 differences is less convincing given the lower PCC. We conjecture that it takes more reliable human ratings to generate better automatic trend predictions.

8.7.4 Conclusions for these experiments

Comparing results from IPMs (Intelligibility Prediction Models) built on Flemish and Dutch acoustic models respectively, we could establish that alignment-based methods are clearly sensitive to the language whereas alignment-free methods are not. Comparing results emerging from IPMs built on different text fragments, we discovered that alignment-based methods are more sensitive to changes in the text material, even though the sensitivity to the text used during IPM training is low.

Our experiments show that by using one single speaker feature set, we were

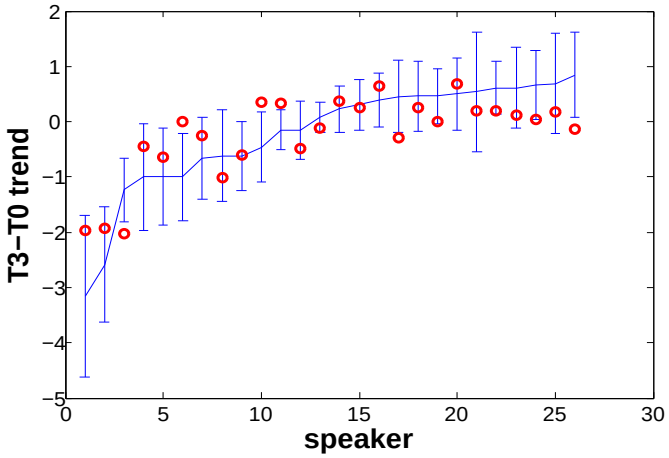

Figure 8.5: Measured and predicted intelligibility trends between T0 and T3 for the speakers exhibiting a clear trend (see text).

unable to create an IPM that is as reliable as a human rater. On the other hand, by combining the Dutch versions of the alignment-based PMF and PLF we already get a human-machine correlation of 0.80, which is only slightly worse than human-based evaluations. Combining alignment-free and alignment-based phone(mic) features leads to a model that can compete with a human rater for comparing one pathological speaker to another. Moreover, the IPM built on these two feature is capable of detecting progress or deterioration of a patient to the same extent humans can.

As the dataset not only contains speech intelligibility ratings but also ratings concerning articulation, voicing etc., future work will focus on the further development of a robust diagnosing system that also offers a more detailed speaker profile concerning articulation, voicing etc. From such a profile one could then retrieve objective and detailed information about the progress of a certain patient in the course of a therapy as well as information which could help determining the right personalized therapy for each patient.

9

Towards a full characterization of pathological speech

9.1 Introduction

In the previous chapters, we showed that the derived feature sets form the base of a simple and robust intelligibility prediction system. We extended our system from a word-based setting, as used in the DIA test, to a more general setting that can achieve a very good prediction of intelligibility starting from running speech. The developed methods even proved to work on another accent (Dutch) and a closely related foreign language (German) and to be robust against changes in the text. In addition, we demonstrated that our method shows close-to-human performance for the tracking of changes in speech intelligibility of one person subjected to a treatment. These achievements clearly show that -using the techniques developed in the current thesis- speech intelligibility can be predicted reliably by a computer and that it can help the speech therapist to judge a patient's (progress in) intelligibility.

However, knowing how intelligible a person is, is not enough to determine the right therapy for this person. The speech therapist also tries to determine the underlying reasons for a low intelligibility by investigating phonetic contrasts, phonation of a sustained vowel, capabilities in terms of diadochokinetic rate, formant transitions and many other possible aspects [12, 95, 121].
In this respect, an additional way of using the DIA test is not only to obtain the intelligibility, but also to derive from phoneme shifts between targeted and perceived phonemes what the underlying articulatory problems might be: perceiving /b/ as

/p/ points to voicing problems, /I/ as /E/ to centralization of the vowels (vowel place problem), etc. For this purpose, the DIA test developed a confusion matrix where all possible phoneme shifts are labeled as correct or incorrect with respect to voicing, manner, place, vowel place, rounding, vowel height or as an addition or omission. For more details regarding this matrix, we refer to the DIA manual [6]. The resulting segmental analysis might contribute to a selection of the right personal therapy [68]. A small study by Van Nuffelen et al [68] on 30 persons with varying speech disorders, judged by nine experienced speech-language pathologists, indicated that the inter-rater agreement at this level is proportional to the speaker's intelligibility. Only for slightly to moderately impaired speech, the segmental analysis is reliable enough to be used as a base for therapeutic decisions.

The COPAS database provides such segmental analyses for many speakers, be it on the basis of the scores of only one human rater. Starting from these analyses of speakers in 121_TM, 121_DIA and 231_DIA, we investigated whether we were able to automate the therapist's analysis. Results on these experiments are discussed in Section 9.2.

The NKI-CCRT-corpus provides another type of analysis. Here, thirteen listeners had to rate different aspects of the 55 patients' speech. Not only intelligibility, but also phonation quality, accent, articulation, speed, voicing and nasality were evaluated. These criteria may not be so fine-grained as the specific phoneme-shifts available in DIA, but they are acknowledged to be important in the assessment of a patient's speech in clinical practice. Since all criteria were rated by 13 listeners, this study enabled us to investigate other phenomena than intelligibility and - equally important - to compare the performance of our models to the inter-rater reliabilities on exactly the same data. Results on this database will be discussed in Section 9.3.

9.2 Predicting articulatory problems using COPAS

9.2.1 Identifying specific phoneme shifts

A first objective was to search for ways to evaluate more than only intelligibility. To this end, we aimed to imitate the segmental analysis which is available for all persons who read the DIA test. Before starting this, we analyzed the perceptual confusion matrix that forms the basis of the perceptual segmental analysis. Such a confusion matrix between targeted and perceived phonemes is derived from the 50 tested phonemes of list A, B and C together: It contains mostly zeros, since no more than 50 entries can be nonzero in this 41×41-matrix (41 equals the number of phonemes distinguished by ASR-ESAT, together with the possibility of an omission). 50 or less than 50 values differ from zero. Since every tested phoneme

appears once in every list, a matrix element describing the perception(s) of a consonant can have values between 0 and 2 (since two lists examine consonants), and a matrix element describing the perception(s) of a vowel can have values between 0 and 1 (since only one list examines vowels).

To have an idea which phoneme confusions were sufficiently present in COPAS, we counted the number of occurrences of every possible confusion. Table 9.1 displays the phoneme confusions occurring at least 40 times over all speakers, together with a possible explanation for the specific confusions.

confusion	explanation
/m/ heard as /n/	very close sounds
/n/ heard as /m/	very close sounds
/G/ heard as /h/	possibly a dialect issue rather than pathological
/d/ heard as /t/	voicing of plosives is a problem in laryngectomy and hearing impairment
/f/ heard as /v/	voicing of fricatives is difficult for laryngectomees and hearing impaired speakers
omission of /h/	voicing problem or dialect issue
/i/ heard as /I/	centralization of vowels occurs in hearing impaired speakers
/I/ heard as /E/	centralization of vowels occurs in hearing impaired speakers or a dialect issue
/A+/ heard as /u/	loss of diphtonguation, vowel height is a difficult dimension for hearing impaired speakers

Table 9.1: Phoneme confusions frequently occurring in the perceptual DIA tests. The right column gives a possible explanation for the confusion.

For these specific phoneme confusions, a score was created for every speaker, indicating the percentage of times the targeted phoneme was confused with the perceived phoneme. This resulted in scores between 0 and 1, which show a discrete distribution with a major peak around 0, indicating that most speakers did not realize that specific error. To model these scores, two possibilities were investigated: comparison with the automatically created confusion matrix and creating a model per confusion via the previously derived feature sets.

9.2.1.1 Comparison of perceptual and automatic confusions

As described in Section 7.2, WAR-ESAT was derived by letting the ASR-ESAT select the perceived phoneme from a list of possibilities, and by comparing it to the target phoneme. This strategy is simply an imitation of the perceptual test, also resulting in an automatically generated phoneme confusion matrix. Therefore, a

direct comparison between the perceptual and the automatic confusion entries was an option.

The results of this comparison are rather disappointing: PCCs on 231_DIA lie between 0.25 and 0.50, which is far too low to be useful in therapy.

Note that ASR-ELIS was used for deriving the WAR as well, but in Chapter 7 we showed that intelligibility predictions from this ASR are clearly inferior to these of ASR-ESAT. Therefore we did not test ASR-ELIS for this even more specific task.

9.2.1.2 Creating confusion models

Building regression models using the previously derived feature sets PLF, PMF and WAR on 231_DIA and even trying the features emerging from running speech (PLF, PMF, ALF-PLF and ALF-PMF) on 121_TM and possible combinations thereof did not yield PCCs higher than 0.40.

9.2.1.3 Conclusion

These results clearly show that we are currently not able to predict specific phoneme confusions as emerging from the perceptual confusion matrix of a single rater. However, literature already proved that this perceptual analysis is not reliable for speakers with a low intelligibility [68]. More importantly, since these phoneme confusion scores are all based on a very small number of examples per phoneme, they are definitely subject to a lot of noise. Obviously, the situation would improve if scores from many human raters were available for all speakers.

9.2.2 Identifying problematic phonological dimensions

Since finding specific phoneme shifts proved to be too hard, we started to "zoom out" on the ratings. Therefore, we tried to identify phonological problem classes. Starting from the perceptual phoneme confusion matrix, we can determine for every patient a set of phonological scores expressing how many voicing, manner, place, vowel place, rounding or vowel height mistakes were made. We can also construct these phonological class scores from the automatic confusion matrix emerging from ASR-ESAT, leading to a feature set CONF-ESAT.

In trying to model the new phonological scores, we constructed regression models using CONF-ESAT and also the previously derived feature sets on 231_DIA. For 121_DIA, we used CONF-ESAT, PLF, PMF and combinations, for 121_TM we used models derived from PLF, PMF, ALF-PLF, ALF-PMF and combinations thereof.

Table 9.2 presents the best five-fold cross validation results for the six phonological class scores.

class	range	set	RMSE	PCC	feature set	learner
voicing	0-16	DIA	2.06	0.70	PLF	ELR
		TM	1.92	0.77	PLF + PMF	ELR
place	0-8	DIA	1.80	0.49	PLF	ELR
		TM	1.81	0.48	PLF	ELR
manner	0-11	DIA	1.96	0.59	PLF	ELR
		TM	1.96	0.61	PLF+ PMF	ELR
v. place	0-6	DIA	0.77	0.55	PLF	SVR-RBF
		TM	0.76	0.54	PLF + PMF	SVR-RBF
rounding	0-12	DIA	1.91	0.56	PLF	SVR-RBF
		TM	1.82	0.57	PLF + ALF-PLF	SVR-RBF
height	0-7	DIA	1.33	0.50	PMF	ELR
		TM	1.26	0.48	PMF	ELR

Table 9.2: Prediction results for phonological classes. v. place denotes vowel place. Range describes the perceptual score range of the phonological class, set indicates from which recording (DIA or TM) the features originate.

Three remarkable conclusions can be drawn from these results. Although the perceptual scores originate from the DIA test, the prediction results based on features derived from the TM can perfectly compete with these results. Secondly, the PLF set seems to be the most informative set as it creates the best models for most of the classes. This is not surprising since the targeted scores are phonological dimensions and the PLFs almost directly point to these dimensions. Thirdly, only the results for voicing yield a PCC higher than 0.70. All other PCCs are much lower than 0.70, making them unusable in clinical practice. Figure 9.1 displays the scatter plot of perceptual voicing errors versus computed voicing errors for 121_TM using the best feature set combination PLF + PMF.

9.2.3 Predicting partial intelligibility scores

Another research line that we followed was to see whether we could predict the intelligibility of list A, B and C separately. This would indicate whether the problem is more consonant- or more vowel-related, and if it is consonant-related, whether the problems are situated in the initial or final consonant. Van Nuffelen [12] shows that the inter-rater agreement for the three subtests varies from an ICC of 0.80 to an ICC of 0.86 for a set of 30 recordings judged by 9 experienced speech therapists. Since we did not have access to these data, we could not calculate the corresponding PCC or RMSE values. However, compared to the ICC of .91 for the total intelligibility score, the inter-rater agreement drops considerably, especially for list B (testing the final consonant) with an ICC of 0.80.

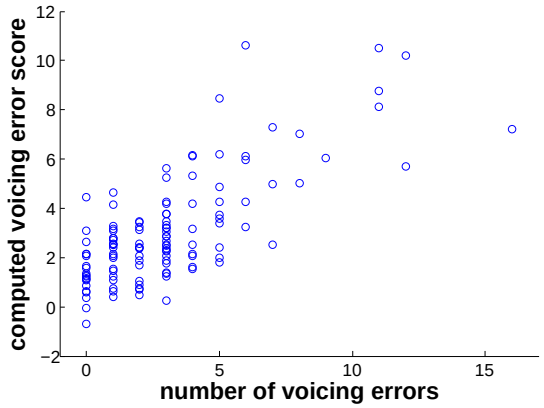

Figure 9.1: Scatter plot of the voicing error prediction.

Adopting the same strategy as before, we trained and evaluated IPMs for list A, B and C separately on 121_TM and 121_DIA. Table 9.3 shows the best prediction results for the three lists.

list	range	rec.	RMSE	PCC	feature set	learner
A	3-19	DIA	2.01	0.77	WAR + PLF + PMF	SVR-RBF
		TM	2.22	0.70	PLF + PMF	ELR
B	3-15	DIA	1.77	0.58	WAR + PLF	ELR
		TM	1.77	0.55	PMF + ALF-PMF	SVR-RBF
C	3-16	DIA	1.59	0.73	WAR + PLF	SVR-RBF
		TM	1.87	0.67	ALF-PLF	ELR

Table 9.3: Prediction results for intelligibility sublists.

While for other experiments results based on running speech and on words were rather similar, here the word-based results are slightly better than the running speech-based results. The reason for this is that for 121_DIA, we have access to WAR, which scores point directly to the word accuracy of list A, B and C.

Apart from this, the PCCs seem to be acceptable for list A and list C, but predicting the intelligibility of final consonants (list B) seems to be more problematic. These results follow the trend of the inter-rater reliability, which was also the lowest for list B.

9.2.4 Displaying important dimensions for pathologies

Since in COPAS we only have little reliable information besides intelligibility, we sought for a way to work around the problem. A possible strategy to learn about a speech pathology and its specific characteristics is to determine the most important dimensions which distinguish it from normal speech. Since the feature sets we created (except for WAR) are closely related to articulatory and phonemic dimensions, a limited number of features might be sufficient to get a detailed characterization of the type and severity of the articulatory problems of a certain speaker.

In order to get evidence in support of this argument, we examined the abilities of all 2-dimensional subspaces of the alignment-based or alignment-free speaker feature spaces to make a visualizable distinction between normal speakers and either hearing impaired speakers, laryngectomees, clefts or dysarthric speakers. Interesting subspaces are then defined as subspaces in which this distinction can be made with high accuracy. Since this implies an exhaustive search in a 100- or 600-dimensional space, we searched for a simple and quick classifier scanning all possible combinations of two dimensions to find the most distinctive dimensions. Therefore, we opted for Linear Discriminative Analysis (LDA). This learner minimizes the within-variance of the class centers of normal speakers and of speakers belonging to one specific pathology, while maximizing the between-class variance according to a linear boundary.

As a baseline experiment, we worked on 231_DIA using the alignment-based features except for WAR. We excluded WAR since these three features would not provide more insight in the core problems underlying the speech pathology. Then, we moved to 121_DIA and 121_TM. On the latter we could compare results obtained with alignment-based and alignment-free features. The results on 231_DIA remain interesting since there are no cleft lip and palate children in 121_DIA. We used a 5-fold CV strategy to validate the LDA models. Results are summarized in Table 9.4.

P	231_DIA	121_DIA	121_TM alignment-based	121_TM alignment-free
H	10	21	21	12
C	26	-	-	-
L	5	8	8	7
D	25	25	30	23

Table 9.4: Classification errors (in %) for discrimination between normal speech and speech of one specific pathology.'P' denotes pathology: H(earing impairment), C(left lip and palate), L(aryngectomy) or D(ysarthria).

This table reveals that alignment-based features derived from running speech

(121_TM) or from the separated words (121_DIA) lead to about equally well performing models. One can thus as well opt for the more natural running speech since also intelligibility can be predicted reliably from this type of speech. A second striking result is that the alignment-free feature set is more discriminative than the alignment-based set.

A last remark concerns the differences between 231_DIA and 121_DIA. While the results are the same for class D, they are better for the other classes (H,L). If the latter improvements are due to the presence of more training data, it is strange that this does not hold for the D model. We did not really search for a reason because the alignment-free feature sets are much better and they lead to about the same results as obtained with the alignment-based features on the large DIA-set.

Figure 9.2 shows a scatter plot of the hearing impaired and the normal speakers in the subspace of "mean maximum probability for nasality" and "mean duration of alveolar segments".

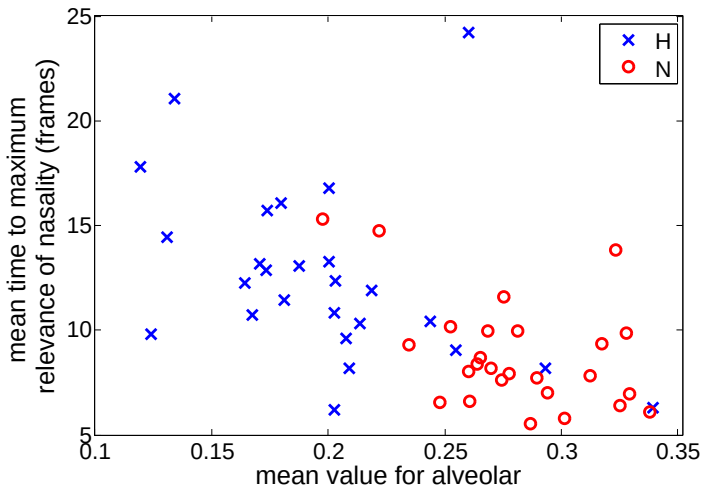

Figure 9.2: Scatter plot of control speakers (N) and hearing impaired speakers (H) in the most discriminative subspace of the speaker feature space.

The figure confirms the findings of [123, 124] that hearing impaired speakers sound hypernasal. Alveolar sounds are also known to be prone to errors [62] since those sounds are not visible and therefore difficult to master by hearing-impaired persons. The depicted feature combination is the best in four of the five folds, and

the second best in the fifth fold.

For the laryngectomees, features for voicing appear to be very discriminating, as could be anticipated. However, less expected is the fact that all the best feature pairs also contain at least one feature concerning turbulence, referring to fricative and plosive sounds. This complies with the fact that in the ASR-based approach (from which the classification accuracy is mentioned in the third column of Table 9.4), we found fricative to be an important feature. Although this needs further investigation, Figure 9.3 seems to support the hypothesis that laryngectomees have difficulties to switch between voiced and unvoiced sounds, in particular between vowels and fricatives or bursts.

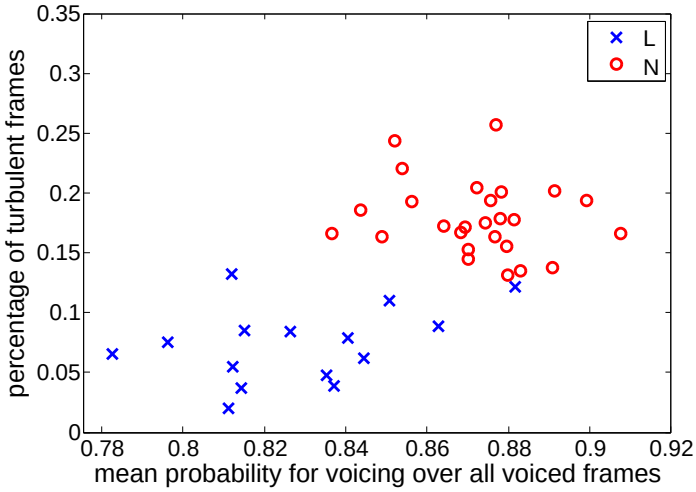

Figure 9.3: Scatter plot of control speakers (N) and laryngectomees (L) in the most discriminative subspace of the speaker feature space.

In the same 5-fold cross-validation experiment, we used LDA to define the optimal linear separation border between the pathological and the control class. Starting from this border, we could determine the distance of each patient to this border as a possible measure of severity of the pathology. Although we believe that this measure can point to problems of speakers in the given dimensions, this distance is not necessarily correlated with the intelligibility score because the main differences in speech characteristics between a pathological speaker and control speakers do not necessarily correspond to the main causes of intelligibility degra-

dation.

9.3 Predicting articulation and phonation problems using NKI-CCRT

The NKI-CCRT dataset, used during our collaboration with NKI-UVAFON, not only contains speech intelligibility ratings but also ratings concerning articulation, phonation quality, etc. of 13 listeners. This allows us to investigate whether automatic methods can be applied for predicting other measures than intelligibility. This would be a first step towards a robust system for automatically creating a full speaker profile concerning articulation, phonation quality etc. From such a profile one could then retrieve objective and detailed information about the progress of a certain patient in the course of a therapy, as well as information which can help determining the right personalized therapy for each patient.

In this section, we investigate the automatic evaluation of perceptual scores for *articulation*, *phonation quality* and *accent*.

Articulation and phonation quality are often reported as correlates and predictors of speech intelligibility [75, 134]. If these dimensions could be reliably evaluated by a machine and combined with an existing model of speech intelligibility [14], one would have a powerful, multi-dimensional evaluation of a speaker. We include the dimension *accent* as speakers have different regional dialects and not all speakers were native Dutch speakers. Articulatory-acoustic variation between speakers can be an effect of regional variation or social background [135] and mother tongue in the case of non-native speakers of Dutch. Although the variable *accent* is not a clinically relevant aspect of speech that requires neither evaluation nor intervention, "accentedness" has been linked to increased listener processing time during signal decoding [136, 137]. By modeling this variable, we envisage that clinicians can take the computed accent score into account when interpreting automatic scores of speech intelligibility: that is, if accent is strongly present caution may be warranted when drawing conclusions on the speaker's computed speech intelligibility score (which is then bound to be underestimated).

In this section, we will develop prediction models for the three perceptual variables and compare model performance to human performance. Subsequently. we will investigate whether these models can track changes in ratings over time.

9.3.1 Perceptual scores in NKI-CCRT

Since the validation corpus and its intelligibility ratings were already discussed in Section 8.7, we limit the description of perceptual ratings to the aspects *articulation*, *phonation quality*, and *accent*.

Articulation Listeners were instructed to evaluate the general precision of vowel and consonant production as compared to normal running speech on a 5-point scale with 1 being 'extremely imprecise' and 5 being 'normal/precise'. Precise articulation was defined as correct manner and place of production and clear coordination between sounds.

Phonation quality Listeners were instructed to evaluate the degree to which phonation quality deviated from what they considered normal. Listeners rated phonation quality on a 5-point scale with 1 being 'very deviant' and 5 being 'normal'.

Accent Listeners were asked to evaluate the weight of the speaker's dialect or accent as compared to standard Dutch (defined as the speech commonly heard on radio and television). Listeners evaluated their perception of accent on a 5-point scale with 1 being 'heavy accent' and 5 'normal/no accent'.

Inter-rater agreement Human performance was calculated as the average RMSE and PCC between the ratings of one individual rater and the group mean score. We use these coefficient averages as the target for model performance. These inter-rater agreements are displayed in Table 9.5.

	Mean	Range
Articulation		
RMSE	0.54	0.36-0.76
PCC	0.75	0.56-0.84
Accent		
RMSE	0.57	0.43-0.91
PCC	0.78	0.65-0.89
Phonation quality		
RMSE	0.56	0.36-0.79
PCC	0.66	0.47-0.78

Table 9.5: Summary of human inter-rater agreement for articulation, accent and phonation quality. Performance is calculated using the RMSE and PCC between the ratings of one individual rater and the group mean score.

9.3.2 Method

We adopt an evaluation strategy that is similar to the one presented in Section 8.7 and [14] where we combine several feature sets to model the reference scores.

Although Table 9.5 shows that the PCC between one rater's score and the average over all 13 perceptual scores can be rather low, the rater causing this low PCC value is a different one for every aspect (articulation, accent, phonation quality). Therefore, there is no single rater we can exclude as being unreliable for all aspects. We will thus continue to use the average over all 13 raters as our reference score. In this way we compare computed scores against human scores as if the computed scores were made by an additional rater.

9.3.2.1 Automatic evaluation

Automatic evaluation of any variable again involves the three stages of processing explained in the previous chapters: (1) front-end analysis of the speech signal, (2) extraction of speaker characteristics and (3) conversion of the feature information to a score by means of a score prediction model.

The acoustic front-end extracts an acoustic parameter vector X_t of MFCCs or mel spectra-based features. Starting from this frame-level information, all the vectors X_t of a speaker are analyzed to derive a number of speaker-level characteristics.

Tested feature sets Here, we consider phonological features (PLFs), alignment-free phonological features (ALF-PLFs), phonemic features (PMFs) and alignment-free phonetic features (ALF-PMF). All feature sets were based on acoustic models trained on Dutch normal speech, as explained in Section 8.7. Since one of the aims of this research line is to predict phonation quality, we also included the AMPEX features which extract pitch-related speaker characteristics. The AMPEX program ([138]) is a voice analysis tool developed by the Digital Speech and Signal Processing research group of ELIS. The program performs a two-step procedure where acoustic information is extracted and is then converted into speaker parameters. The AMPEX version used in this paper provides eight speaker features that can be divided into voicing-related parameters (e.g. the percentage of speech frames classified as voiced) and pitch-related parameters (e.g. average jitter in voiced frames). See [139, 140] for further information.

Prediction model We utilize the same strategy as in Section 8.7 in which prediction models are trained and evaluated using a 5-fold cross validation strategy using ensemble linear regression models. In addition to the five individual features (PLF, PMF, ALF-PLF, PLF-PMF, AMPEX), we also tested four feature combinations: full forced alignment using both phonemic and phonological features (PLF+PMF), combined phonological sets (PLF+ALF-PLF), combined phonemic sets (PMF+ALF-PMF) and combined phonological features with AMPEX (PLF+ALF-PLF+AMPEX).

9.3.3 Predicting articulation, accent and phonation quality

In a first study we investigated which individual feature sets or combined feature sets produce the best prediction models for the three investigated variables. Performances of all created models are listed in Table 9.6 together with model target human rater performance for each variable. In cases where the RMSE is the same between models, we use the PCC to differentiate between model performance.

	Articulation		Accent		Phonation	
Models	RMS	PCC	RMS	PCC	RMS	PCC
Inter-rater agreement	*0.54*	*0.75*	*0.57*	*0.78*	*0.56*	*0.66*
PLF	0.44	0.75	0.56	0.72	0.55	0.39
PMF	0.45	0.74	0.59	0.67	0.59	0.24
ALF-PLF	0.51	0.66	0.68	0.55	0.58	0.33
ALF-PMF	0.45	0.75	0.65	0.64	0.60	0.23
AMPEX	0.66	0.24	0.82	0.30	0.55	0.43
PLF+PMF	0.44	0.75	0.56	0.71	0.54	0.42
PLF+ALF-PLF	0.44	0.78	0.55	0.74	0.53	0.47
PMF+ALF-PMF	**0.42**	**0.80**	**0.54**	**0.77**	0.58	0.27
PLF+ALF-PLF+AMPEX	0.44	0.78	0.56	0.71	**0.46**	**0.62**

Table 9.6: Performance of prediction models using different feature sets for the variables articulation, accent and phonation quality. The best performance is indicated in bold. Results differing significantly ($p < 0.05$) from the best result are underlined. Phonation stands for phonation quality. RMS denotes RMSE.

9.3.3.1 Articulation prediction models

Four of the five single feature sets yield a performance which is competitive with that of the average human listener. This could be partly due to the fact that human ratings are discrete values which will tend to agree less well with an average (continuous) score than continuous automatic scores. Almost all single feature sets yield similar results, except for AMPEX, which is obviously unable to predict articulation as could be expected since it contains no features related to place and manner of articulation. Combining feature models results in a small but significant improvement compared to individual features. The combination of ALF-PMF and PMF achieves the best performance with a RMSE 0.42, which is significantly better than all prediction models based on one feature set. It is however not significantly better than combining both phonological feature sets ALF-PLF and PLF. Adding AMPEX to this last combination does not improve on the results.

Figure 9.4 displays a scatterplot of computed and observed mean scores. This

plot confirms the earlier findings of Van Nuffelen et al. [68] that articulatory problems are difficult to analyze in speakers with a low intelligibility. In both [12] and our study, this may be due to the low prevalence of speakers with low perceptual scores.

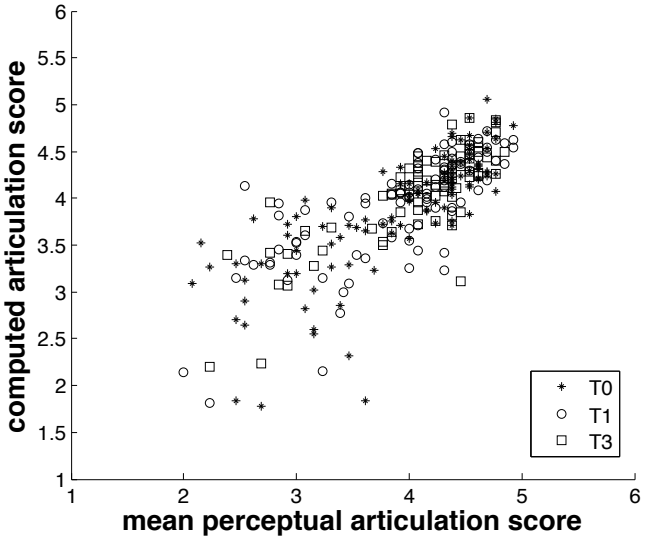

Figure 9.4: Correlation between perceptual and computed scores for articulation.

Selected features in best performing models As we utilize an ensemble linear regression approach to develop the prediction models, we can inspect how often certain features are selected in the smaller models to gain insight into the important dimensions for modeling articulation performance of the speakers.

Table 9.7 presents the features selected at least ten times in the most predictive model based on ALF-PMF and PMF. Features related to the vowels /@/, /i/, /A/ and /A+/ and consonants /s/, /l/ and /n/, /d/ are selected.

Feature	/i/	/s/	/n/^{a}	/l/	/l/^{a}	/A/	/u/	/s/^{a}	/d/	/@/	/A+/^{b}
Frequency	44	36	28	25	23	22	16	16	15	15	11

Table 9.7: Features selected and their frequency in best performing articulation models. ^{a}percentage of frames in which x was recognized. ^{b}mean evidence of feature x over all frames where feature x was recognized.

As a whole, these features appear to represent the diagonal of the vowel trapezium and production of anterior Dutch lingual consonants, suggesting that range

and precision of tongue movement are important aspects in modeling perceptual scores of articulation. Difficulty with production of anterior lingual consonants, particularly alveolar fricatives, for speakers treated non-surgically for cancer of the head and neck has also been reported [141].

The features selected for the articulation prediction model also overlap with several of the features selected for speech intelligibility prediction models (see Section 8.7 and [14]), for instance features of vowels from the diagonal of the vowel trapezium (/i/, /@/, /A/) and selection of features related to /n/ production. The similarities between the features in the models for articulation and speech intelligibility are not surprising given that intelligibility is considered a very important factor of articulation.

9.3.3.2 Accent prediction models

For accent prediction, the same trends as for articulation are observed, but the individual feature set models are now marginally worse than the average human listener. The individual feature sets, except AMPEX, show again a rather similar performance, be it that PLF stands out a little more now in a positive sense (leading to not significantly different results from the best model) and ALF-PLF a little more in the negative sense. Although the improvements are not statistically significant yet, the data seem to indicate that PMF and ALF-PMF exhibit at least some complementarity that can be exploited. Striking is that combining PMF with PLF is not helping at all.

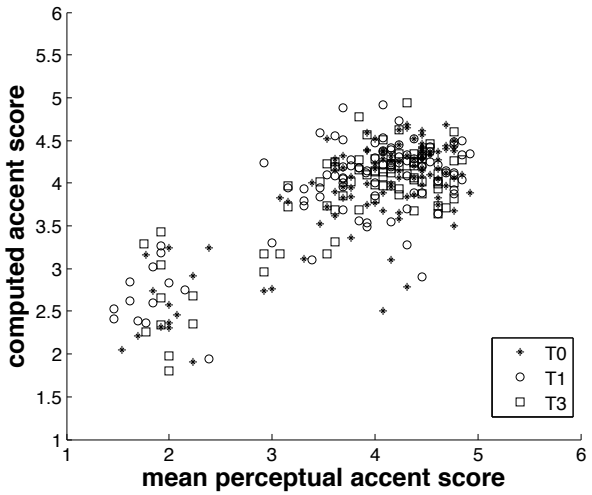

Figure 9.5: Correlation between perceptual and computed scores for accent.

Selected features in best performing models The combined phone(mic) accent prediction model selects five features related to vowels /A/, /Y+/, /a/, /y/, and /E/ and three consonant features /h/, /n/ and /z/ (see Table 9.8). Unlike the articulation prediction model, accent features do not focus on differentiation of vowels in the trapezium, but rather seem to suggest that differentiations in high and low central vowels are important acoustic indicators of speaker accent. Four of these vowels (/A/, /Y+/, /a/ and /y/) are reported as being difficult for non-native Dutch speakers [142] and may explain their inclusion as selected features.

Selection of the phonemic feature for the glottal fricative /h/ can be explained in two ways: as a phoneme non-native Dutch speakers often produce incorrectly [142] and/or as a phoneme with variable realizations in which speakers may not apply the expected /h/ omission rule for unstressed syllables [29].

Feature	/Y+/	/A/	/n/	/y/[a]	/z/[b]	/a/[c]	/E/[c]	/h/
Frequency	41	38	25	19	15	9	9	9

Table 9.8: Features selected and their frequency in best performing accent models. [a]percentage of frames in which x was recognized. [b]standard deviation of probability of x. [c]mean evidence of feature x over all frames where feature x was recognized.

9.3.3.3 Phonation quality

The RMSE results for predicting phonation quality displayed in Table 9.6 again indicate that all models based on one feature set perform about equally bad: although single-feature they perform in the range of human raters, none of the models outperforms the average human rater. Unlike the single feature models for articulation and accent, the AMPEX model has a slight performance advantage over the phonological and phonemic models. Combinations of one feature set plus AMPEX were also tested but did not lead to significant improvements. Adding AMPEX features to the best combination of two feature sets, namely PLF+ALF-PLF, does however lead to a significant improvement. The performance of that combination seems to be at the level of human performance (somewhat lower RMSE, somewhat lower PCC). Figure 9.6 displays a scatter plot of computed and observed mean scores for this top-performing model. Striking is that the prediction fails when the phonation is really bad. This may have to do with the fact that the pitch and voiced/unvoiced detector that is at the basis of the AMPEX features is not reliable anymore for recordings with low voice quality.

The strong performance of the AMPEX features is not surprising as these features are designed to analyze phonation-specific aspects of speech. Previous work has already reported good correlations between AMPEX scores and perceptual

judgments of overall phonation quality [139]. The equally strong performance of the PLF and the AMPEX feature sets and the fact that their combination leads to a higher performance suggests that the two feature sets constitute two partly complementary views on voice quality.

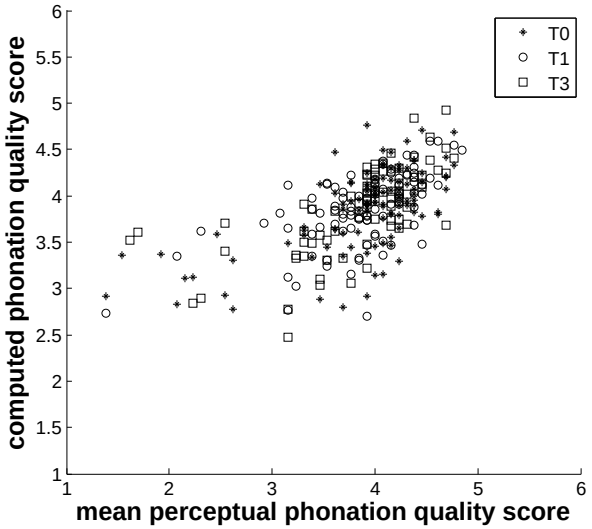

Figure 9.6: Correlation between perceptual and computed scores for phonation quality.

Features selected in the best models The PLF + ALF-PLF + AMPEX model selected 10 features (2 AMPEX, 6 ALF-PLF and 2 PLF, see Table 9.9 for details). Most features can be related to the rate of movement between extremes in lip rounding (round vs. open), tongue position (back vs. front) and phonation control. Other features point to the shape of the oral cavity (features related to lip rounding and tongue position) and inclusion of the nasal cavity (features related to nasality and 'trill' [1]) as a resonator can cause changes to the acoustic spectrum. The model features can also be understood in terms of the tumor location and treatment on the one hand and the source-filter model on the other hand. We would expect that a tumor at the level of the larynx would has effect on phonation and treatment with radiotherapy and chemotherapy causes changes to tissue structures around the tumor site. Likewise, we would expect that tumors and tumor treatment for tumors in the nasopharynx or oropharynx has an effect on the ability to use the oral and nasal cavity as a filter. The features selected in the prediction models may indicate that listeners assess phonation quality on the basis of phonation and resonance information.

[1] Dutch /r/ is highly variable and the trill variant can be produced as an alveolar or uvular trill [29]

Feature	AVE_A	PVS_A	PVF	vowel	unvoiced-voiced[a]
Frequency	35	28	22	22	19
Feature	nasality[b]	front-back[c]	trill	vowel nas.[d]	spread-round[e]
Frequency	12	12	11	11	10

Table 9.9: Features selected and their frequency in best performing phonation quality prediction models. AVE = average voicing evidence in voiced frames (AMPEX feature). PVS = percentage of speech frames classified as voiced (AMPEX feature). PVF = percentage of voice frames. [a]mean time needed to go from unvoiced to voiced. [b]mean minimum value for relevance of consonant nasality. [c]mean time needed to go from relevant to not relevant in front-back dimension. [d]standard deviation of vowel nasality probability in frames in which vowel nasality is present. [e]mean time needed to go from spread lips to rounded lips.

9.3.4 Tracking trends over time

Treatment for cancer of the head and the neck can have negative effects on phonation quality and speech production [129, 141, 143]. We investigated whether the models which could predict phonation quality and articulation for a population of patients can also track changes over time of these variables for a single patient. We do not include the dimension *accent* as we do not expect this aspect to change between evaluation moments. For predicting articulation we use the best combined phonemic model (PMF+ALF-PMF). For predicting phonation quality we use the best model supplied with a combination of phonological and AMPEX features (PLF+ALF-PLF+AMPEX). We let the models compute scores for each speaker at the different times T0, T1 and T3 and from these scores we derive differences between T0 and T1, T1 and T3 and T0 and T3. Using the same methodology as in Section 8.7, we estimate human performance by computing the RMSE and PCC between the differences retrieved from the scores of one individual rater and the mean over all raters. Similarly, we estimate model performance by computing the RMSE and PCC between the automatic score differences and the mean score differences. The results of all this lead to Table 9.10.

Since a lot of speakers did not exhibit any trend, we created per variable and per time-pair an evaluation set consisting of all patients for which the human raters seemed to agree on the presence and direction of the trend, just like we did in Section 8.7. The results for the evaluation sets are listed in Table 9.11.

The data in Table 9.11 show that the model predictions are better than those of the average human rater (lower RMSE and higher PCC) in case of T0-T3, but not in the other cases involving T1. In order to establish a reason for this we first of all inspected the histograms of the human score-differences between T0-T1, T0-T3 and T1-T3 (see Figure 9.7). Apparently, there is no significant difference between

		Perceptual		Computed	
Evaluation	N	RMSE (range)	PCC (range)	RMSE	PCC
Articulation					
T0-T1	93	0.71 (0.49-1.05)	0.46 (0.30-0.64)	0.45	0.35
T1-T3	76	0.73 (0.51-1.11)	0.49 (0.38-0.60)	0.45	0.19
T0-T3	76	0.67 (0.42-0.99)	0.48 (0.20-0.66)	**0.36**	**0.55**
Phonation quality					
T0-T1	93	0.74 (0.51-1.05)	0.61 (0.50-0.77)	0.49	0.53
T1-T3	76	0.74 (0.45-0.98)	0.49 (0.37-0.66)	0.52	0.22
T0-T3	76	0.72 (0.45-1.05)	0.55 (0.37-0.74)	**0.40**	**0.59**

Table 9.10: Overall performance for computing changes in articulation and phonation quality between two evaluation moments. Values in bold highlight computed results that are better than the corresponding perceptual ones. N is to the number of recordings included in the comparison.

		Perceptual		Computed	
Evaluation	N	RMSE (range)	PCC (range)	RMSE	PCC
Articulation					
T0-T1	19	0.58 (0.37-1.04)	0.69 (0.43-0.92)	0.58	0.22
T1-T3	17	0.62 (0.38-0.89)	0.74 (0.55-0.85)	0.69	0.04
T0-T3	21	0.57 (0.42-0.77)	0.70 (0.28-0.84)	**0.40**	**0.72**
Phonation quality					
T0-T1	27	0.73 (0.53-0.97)	0.77 (0.61-0.89)	0.60	0.76
T1-T3	18	0.60 (0.46-0.86)	0.75 (0.61-0.89)	0.61	0.45
T0-T3	17	0.64 (0.42-1.01)	0.79 (0.62-0.92)	**0.45**	**0.86**

Table 9.11: Performance for computing changes in articulation and phonation quality between two evaluation moments, but only in cases where human raters agree on the direction of the trend. Values in bold highlight computed results that are better than the corresponding perceptual ones. N is the number of recordings included in the comparison.

the shapes of these histograms that correlates with the differences in performances emerging from Table 9.11.

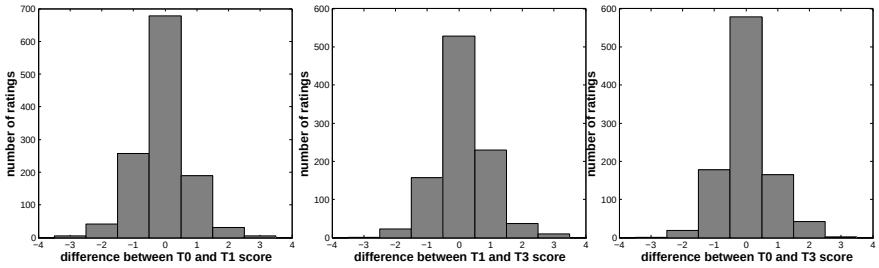

Figure 9.7: Histograms of the human score-differences for articulation between different evaluation moments.

Another hypothesis for the bad trend predictions in comparisons involving T1 was that the score predictions at T1 were worse than at other moments. To verify this we have also computed the performances of the articulation and phonation quality prediction model on all available recordings made at one evaluation moment. From Table 9.12 it follows that both the perceptual and the automatic scores are roughly equally reliable for T0 and T1, and less reliable for T3.

	Target		Computed	
Evaluation	RMSE (range)	PCC (range)	RMSE	PCC
Articulation				
T0	0.53 (0.32-0.70)	0.79 (0.57-0.89)	**0.45**	**0.83**
T1	0.60 (0.40-0.80)	0.73 (0.55-0.85)	**0.43**	**0.76**
T3	0.50 (0.33-0.79)	0.67 (0.49-0.81)	**0.36**	**0.75**
Phonation quality				
T0	0.56 (0.32-0.85)	0.65 (0.44-0.84)	**0.50**	0.63
T1	0.56 (0.38-0.72)	0.62 (0.47-0.70)	**0.43**	0.60
T3	0.61 (0.41-0.79)	0.53 (0.36-0.62)	**0.60**	0.45

Table 9.12: Model performance for recordings at each evaluation moment. Values in bold highlight performance measures that meet or exceed target level. T0 = pre-CCRT. T1 = 10-weeks post CCRT. T3 = 12-months post CCRT. N refers to the number of recordings included in the comparison.

A third hypothesis was that there were maybe more, usually unreliable low

scores (≤ 3) from which to derive the trends in the comparisons involving T1. The percentages of low articulation scores were 7% for T0-T3, 11% for T0-T1 and 12% for T1-T3. The number of low phonation quality scores were 6%, 9% and 15% respectively. This seems to suggest that less low scores may lead to better trend predictions.

Figure 9.8 shows some plots which illustrate that almost all automatic trend predictions fall within the limits of the error bars, showing that they would be acceptable human ratings. However, the mean human trends are close to zero, with only a few speakers showing a clear difference higher than 1 or lower than -1. This means that the PCC will mainly be determined by these few speakers: it will be high or not if the trends observed for these few speakers are well predicted by the model or not. Moreover, the high RMSE values between one rater's scores and the mean perceptual ratings suggest that the relatively poor performance of the computed scores to track change over time is due to the variability in human scores at the evaluation moments and is not due to a model's lack of sensitivity to change.

As a last experiment we performed an analysis of the model and rater capacities to perform a three-fold trend classification: negative change ($\leq$ -0.5), no change, and positive change (≥ 0.5). Rather than evaluating all speakers we limit our analysis again to the recordings selected above for which all raters "agreed" about the differences for the considered time-pair. This way, we retain 57 derived scores for articulation (29 for T0-T1; 17 for T1-T3; 21 for T0-T3) and 62 for phonation quality (27 for T0-T1; 18 for T1-T3; 17 for T0-T3). Using our best articulation and phonation quality prediction models, trend classification accuracies are 72% and 65% respectively. As can be seen in Table 9.13, all disagreements are within one category: there is no case where the observed trend is positive and the predicted trend is negative or the observed trend is negative and the predicted trend is positive.

Articulation				Phonation quality			
Predicted	Observed			Predicted	Observed		
	-	±	+		-	±	+
-	3	0	0	-	7	3	0
±	7	35	8	±	14	26	4
+	0	1	3	+	0	1	7

Table 9.13: Contingency table between predicted and observed articulation and phonation quality trends for negative change (-), no change (±) or positive change (+).

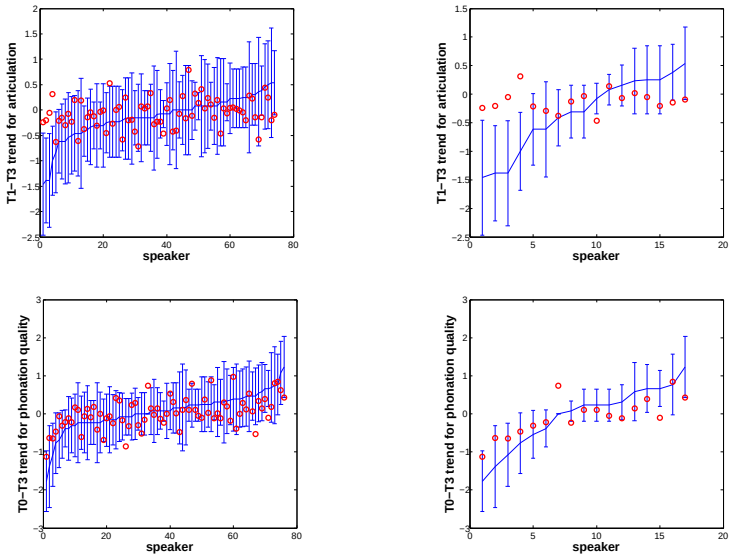

Figure 9.8: Trends derived from human ratings (curves connect the means, vertical lines indicate the standard deviation over all raters) and trends predicted by the models (circles). On the left one sees results for all speakers, on the right only for the selected speakers. On top one sees the results for articulation differences between T1 and T3 (the lowest PCC in Table 9.11). At the bottom one finds the results for phonation quality differences between T0 and T3 (the highest PCC in Table 9.11.

9.3.5 Conclusions for this section

The prediction models presented in this section have been developed on a group of Dutch speakers with cancer of the head and neck which were treated by means of CCRT. The aims of this study were (1) to investigate whether mean perceptual scores of articulation, accent and phonation quality can be automatically evaluated at a level comparable with that of human raters and (2) to investigate if these models can track changes in perceptual scores of one patient over time. This study is unique as the prediction model is based on perceptual scores of a relatively large number of listeners, and these listeners are also a group of semi-professionals.

We have shown that prediction models combining forced alignment and alignment-free speaker feature sets (and AMPEX features for phonation quality) yield correlations between perceptual and computed scores that are within the range of human performance. In the case of articulation prediction, the model performance

even exceeds the performance level of the average human rater. However, articulation and phonation quality prediction models attain varying levels of success when tracking change over time within a single speaker. There seems to be some evidence that a part of that variability may stem from the difficulty human raters experience when rating bad voices. If the low scores are less reliable, the trends derived thereof may be even more unreliable and unreliable targets during model training will of course lead to unreliable models as well. We envisage that future work will focus on getting more reliable perceptual trend data, e.g. by normalizing perceptual scores in order to reduce skew in the perceptual data.

It has been shown however that a categorization of the trends in three classes (positive change, no change, negative change) can be achieved at human performance level. In our experiments, it never happened that a positive perceptual trend was classified as being negative and vice versa.

10 The DIA tool

10.1 Introduction

Since this research led to a reliable system for automatic and therefore objective intelligibility assessment, a logical next step is to make this system useful in clinical practice. Software packages to measure intelligibility have been developed before. In [4] one describes a tool for evaluating speech of English patients suffering from dysarthria. In [3], one describes a system doing the same for German laryngectomees and children with cleft lip and/or palate.
The former tool computes the so-called goodness of fit of the alignment between the uttered speech and the target speech, the latter uses the word accuracy of an automatic speech recognition of the uttered speech.

Here, we present the Dutch Intelligibility Assessment (DIA) tool for assisting speech therapists in assessing patients suffering from pathological speech. The tool is based on the methodology described in Chapters 7 and 8. The method underlying the DIA tool extracts phonemic and phonological features from automatic speech alignment on the basis of acoustic models that were trained on normal speech. Based on those features, intelligibility is predicted by means of a compact model that was trained on pathological speech samples. The experimental evaluation of the system has shown standard deviations between perceived and computed intelligibilities that are lower than 8%. This was considered a sufficiently strong result to convince speech therapists in Flanders and the Netherlands to use an automated DIA test that could be made freely available to them via the

internet.

10.2 The DIA Tool for adults

While the perceptual DIA test only uses the 50 tested phonemes, the computerized version takes every phoneme of the 50 words into account. All uttered speech is lined up (forced alignment) against the target words by two ASRs. These two alignments result in two feature sets: phonemic features (PMFs) coming from one ASR and phonological features (PLFs) coming from the other. These feature sets are then used by a simple regression model to predict the intelligibility of the speaker.

Different models have been designed: one general model, as well as pathology-specific models for people with hearing impairment, dysarthria, laryngectomy and for children with cleft lip and palate. Although we recently developed more accurate models by adding extra features and by using SVR instead of ELR, these models have not yet been included in the DIA tool so far, but they will be added soon.

As shown in Chapter 7, the reliability of the predicted scores matches that of a human rater. Therefore, the DIA tool offers an objective and less time-consuming way to administer the test.

Our purpose was to design a user-friendly tool which does not require a complex setup to administer the test. To use the DIA tool, the user only needs a PC or laptop with a web browser, a head set , a sound card and an up-to-date Java runtime environment. The tool works in a client/server environment and can be used both in online or offline mode.

Once a user has an account, he/she can add and edit patients. As we respect the privacy of the patients, every user can only access the data and recordings of his/her own patients. Once a patient is added, the user can start the test of this patient. We advice to do a microphone test first, to be assured that the recording quality is good enough and the microphone is in the right position (e.g. not too close to the mouth). When starting the test, a sequence of words is presented to the patient (see Figure 10.1). Each word is recorded in a separate .wav file. If for any reason the therapist wants the patient to repeat a word, he can achieve this. Only the last recording is stored for subsequent analysis.

When the recording is finished, the speech therapist can perform a perceptual analysis by listening to every word and by filling in the missing phoneme as shown in Figure 10.2. This results in a perceptual score and a report of the segmental analysis displaying the nature of the errors, e.g. wrong place/manner of articulation, as described in Chapter 9 and in the originally developed DIA perceptual test manual [6]. Every recording can also be judged by several listeners, so as to get a better perceptual score.

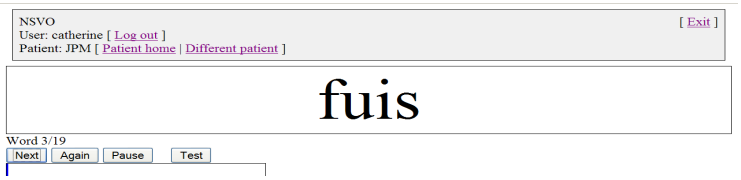

Figure 10.1: Screenshot of the recording environment of the tool. The patient is presented the word which has to be read. Under the buttons "next" etc. a volume control bar displays the volume of the recorded word, with a blue zone for small volume, green for medium and red for high volume.

The user can also run an automatic analysis though. This step results in an objective intelligibility score, as well as a number of visual representations of the analysis results as shown in Figure 10.3. These images show the speech profile of the current patient, compared to normal speakers, as well as to a number of well-defined pathologies according to Section 9.2.4.

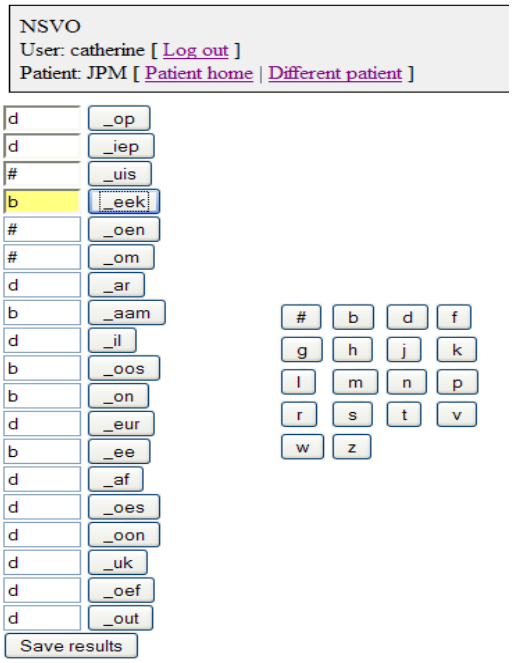

Figure 10.2: Perceptual analysis of the recordings. When clicking on the button, the corresponding .wav file is played, and the listener can fill in the missing part.

To validate the tool, a master student recorded 33 laryngectomees, 19 hearing impaired, and 9 dysarthric patients. The recording settings were not always

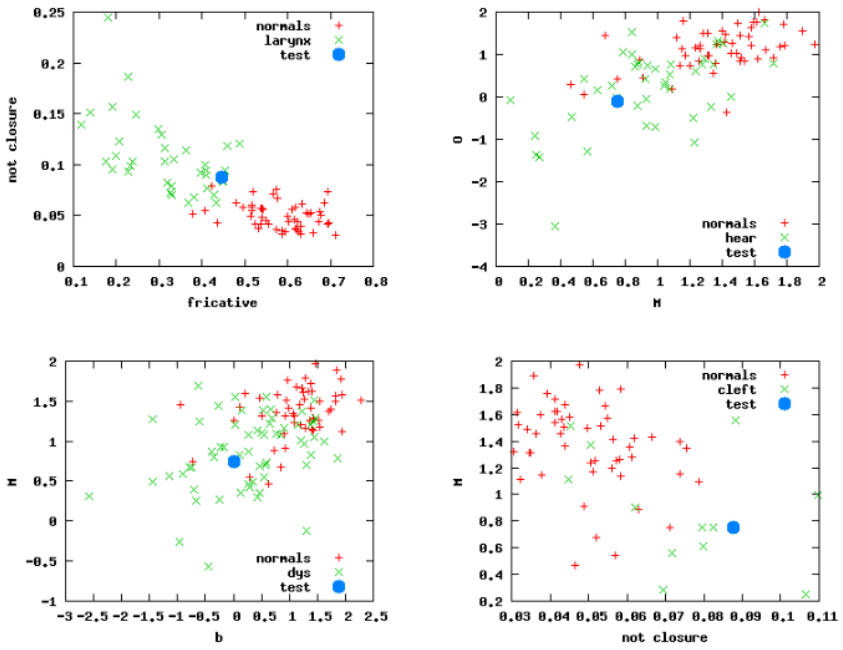

Figure 10.3: Comparison of patient with speaker groups. For every pathology, the current speaker (denoted with 'test') is compared to the control speakers and the pathology according to the two most distinctive dimensions.

ideal and sometimes a lot of background noise could be noticed. Every patient performed the test, which was recorded using our DIA tool. Apart from the objective score calculation, the subjective evaluation of the speech intelligibility was performed by two professional listeners. The inter-rater agreement between the two listeners was measured using the Pearson correlation coefficient between their scores. The measured PCC was as high as 94%. The Pearson correlation between the mean of the listeners' scores and the objective scores reached 90%, which is almost as good [122]. Moreover, the DIA tool was described as a user-friendly and time-saving device.

10.3 The DIA tool for children

Since the word-based DIA tool proved to produce reliable intelligibility scores, the creators of the perceptual DIA test (Marc de Bodt et al, [6]) wondered whether the DIA tool could evolve towards a tool that employs more natural words (since only human listeners are then biased by their language knowledge but the computer

stays unbiased) and towards a tool that could be used by children as well. For that reason a new list of existing CVC words was developed and a visualization of these words for children was included as an option.

The new word list was developed by a master's thesis student. This list consisted of 42 easily visualizable CVC-words [144]. These words were carefully selected using the 'Streeflijst Woordenschat voor Zesjarigen', which is a list containing the words which a six-year old child should know in its passive vocabulary. Moreover, the words were chosen in such a way that - like in the original DIA test - each consonant appears at least twice (in initial and final position) and all vowels and diphtongs also appeared at least twice. The DIA tool includes this new test as the D-list. All words are displayed as in the original test, but now they are also illustrated with pictures from the same sizes and a uniform white background. See Figure 10.4 for an example. Since the D-list is constructed to examine the same phoneme set as the DIA test for adults, it can be analyzed using the same automatic strategy.

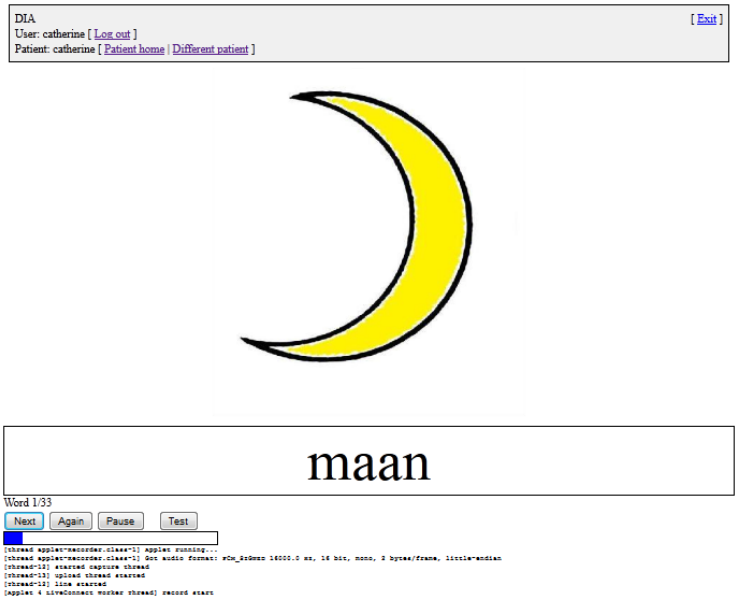

Figure 10.4: Screenshot of the recording of the pictorial DIA test.

In a preliminary study, 14 dysarthric patients read the original DIA test words (50 words) as well as newly included D-list. An automatic intelligibility score was derived from both sets of recorded words. Comparing both automatic scores led to a PCC of 0.78. In general, higher scores emerge for the new test items. This needs to be investigated in more detail. Obviously, the differences can be caused by the fact that fewer hesitations are made on existing words, less background noise was

perceived and the fact that the words from DIA were first recorded so that the patients were already used to the way of testing when arriving at the D-list [144].

10.4 Towards a DIA tool with running speech

Since we already showed the ability to predict phoneme intelligibility rather well starting from running speech, another alternative for non-existing words is using the paragraph 'Papa en Marloes' as test material. Therefore, we plan to include an extra test in the DIA tool which will be available for automatic testing only. While in the case of word tests only one word at the time is presented, the running speech test will be displayed sentence by sentence. Using ALF-PLFs and ALF-PMFs, the patient will be depicted against reference pathologies as described in Section 9.2.4.

11

Conclusion and future prospects

11.1 Conclusions

Automatic analysis of pathological speech is a challenging and very interesting research topic. Knowledge of the complex human speech production system and possible flaws in it, basics of phonetics and speech analysis have to be combined in the search for suitable measures which can help speech therapists to objectively measure severity of a pathology and to gain insight in the underlying causes.

In Chapter 1, I started with a motivation for this work. Intelligibility - a popular measure for severity of pathological speech - was introduced, as this comprised the main part of my research. Chapters 2 and 3 dealt with the medical side of this interdisciplinary work, namely the human speech production and some speech disorders. The current state-of-the-art intelligibility assessment techniques of pathological speech were described in Chapter 4. Most of these tests are based on perceptual measurements, which are subjective in nature. This opens the door for objective, automatic intelligibility assessment tools.

Until now, there are only a few tools available. This striking lack of automatic measures is partly due to the fact that the golden standard in intelligibility testing still is the perceptual test because there is still insufficient confidence in the abilities of computer models. Therefore, my main goal was to develop an automatic intelligibility assessment model for Flemish, based on the perceptual Flemish DIA test described in Chapter 4. In order to create the right model, the right machine learning technique has to be adopted. Basics of machine learning techniques were

therefore treated in Chapter 5. Since building a model requires a database, the used Flemish databases were described in Chapter 6.

Chapters 7 to 10 describe the main part of my research, covering all strategies for developing good speech analysis models. Chapter 7 describes the first steps toward an automatic assessment of phoneme intelligibility by automating the perceptual DIA. It is shown that alignment-based methods combining phonemic and phonological features yield a correlation between the subjective (human) scores and the objective (computed) scores of about 0.80 for a general model and up to 0.96 for a pathology specific model.

The correlations for general and specific pathologies compete with the inter-rater agreements measured for the perceptual DIA. The feature set with the most predictive power was the phonological feature set, describing speech in terms of its articulatory aspects. By investigating which exact features are important, we discovered that all features frequently selected by the intelligibility prediction models can be linked to specific articulatory deficits of pathological speakers.

Since the DIA is based on isolated words, half of the time being nonsense words, it lacks naturalness, leading to hesitations and reading errors. Therefore, I searched for a way to circumvent these problems by using meaningful running speech as the basis of a new intelligibility test. Using again a combination of phonemic and phonological features, now based on forced alignment of speech with sentences instead of words, I proved in Chapter 8 that phoneme intelligibility scores of the DIA can also be reliably predicted from running speech recordings. Still, forced alignment methods on a sentence level can fail in the case of reading errors and hesitations. Therefore I developed a set of alignment-free phonemic and phonological features. While the alignment-free phonological feature set does not need any information at all about the read text, the alignment-free phonemic feature set relies on phoneme frequencies derived from the text to compare the utterance of the speaker against an expected distribution of phonemes. The proposed features have been proven to work in a text-independent scenario and are usable for different languages. Again, the features selected by the model could be linked to specific problems in pathological speakers.

Using a combination of alignment-based and alignment-free features on the NKI-CCRT corpus, I could develop a model for intelligibility which was as reliable as a human listener. Moreover, it is capable of tracking changes in intelligibility over time to the same extent a human listener can.

Having shown that intelligibility can be predicted in a reliable way by using phonemic and phonological (alignment-free) features, I moved on to extract more detailed information quantifying the underlying speech deficits of a tested speaker. In Chapter 9, I first explored some ideas using the COPAS corpus, but this corpus only provides intelligibility scores and no detailed information about specific deficits related to e.g. nasality, phonation etc. Moreover, since only one

rater judged all samples, and since every phoneme shift could occur only a small number of times, the segmental analysis appeared to be reliable only for patients with high intelligibility scores.

Although a segmental analysis of the articulatory problems proved to be impossible, I did discover that the alignment-free feature sets can clearly distinguish a specific type of pathology from normal speech in a two-dimensional subspace that can be identified automatically, and that can easily be visualized. In addition, the features of the best subspace can be linked to the specific pathologies.

Using the NKI-CCRT database, containing ratings for a.o. phonation, articulation and accent by 13 human listeners, I could prove that the feature sets I developed in Chapters 7 and 8 can predict these aspects of speech to the same extent as humans can. It is clear though that, once more specific speech aspects than overall intelligibility are investigated, human inter-rater agreement drops significantly, and so does the automatic prediction of these aspects.

My research resulted in the development of the online DIA tool, which is described in Chapter 10. This user-friendly tool is made freely available via the web. To use it, the user only needs a PC or laptop with a web browser, a head set and sound card, and an up-to-date Java runtime environment. The tool works in a client/server environment and can be used both in an online and an offline mode. The isolated word test can be recorded and automatically analyzed, leading to a report describing the patient's current intelligibility and position against other pathological speakers. Perceptual analysis of the isolated word test is also possible. Soon, the RSI test I developed in Chapter 8 will become available as well.

11.2 Future prospects

At the end of this dissertation, I would like to stress some strenghts, weaknesses, threats and opportunies of using an automatic analysis of pathological speech as presented in this work.

A clear advantage of an automatic over a perceptual analysis is its objective nature. It is insensitive to contextual information and unbiased by knowledge about the patient, his/her pathology and the used text. Moreover, in favourable circumstances my methods have proven to reach about the same reliability as a human rater and they are less time-consuming than the perceptual test. The DIA tool can therefore be considered as an extra "objective" listener that can quickly and reliably measure intelligibility and position the patient against some reference pathological speakers. It can help the therapist in screening patients. It can however never be used as a replacement for a thorough examination of the patient by a speech therapist. Rather than replace the speech therapist, the tool should be re-

garded as an objective, complimentary asset, used to enrich the therapist's toolbox.

Until now, an important issue concerns the robustness against changes in background noise, room acoustics and microphone characteristics, which can be significant in recordings made in a clinical environment. Therefore, there is a need for more robust acoustic features in the front-end and for speaker detection techniques to separate the speaker from other voices in the background, including that of the therapist giving hints to the speaker. The latter is especially true when one wants to test children.

From field tests of the DIA tool, I learned that it would also be interesting to incorporate automatic overflow and underflow detectors which ask to re-record an utterance that was tagged as inappropriate. Clear instructions to the speech therapist concerning the way of positioning the head set microphone and a manual verification of the recordings could also enhance the quality of the recordings and thus of the resulting analysis. More generally, the risk of improper use of the tool, leading to wrong results and wrong conclusions, should be avoided by giving a course to the speech therapist in how to obtain the best results and how to interpret these results (know about the fault margin).

Another issue concerns the monitoring of trends. Although I proved that an automatic analysis can detect progress or deterioration in one patient's speech according to independent perceptual evaluations of the sessions, it still needs to be verified whether it can also predict the outcomes of an experiment in which differences between sessions are scored. Likewise, the models for articulation, phonation and accent are designed specifically for patients treated for head and neck cancer. Patients with a different pathology (e.g. cleft lip/palate or hearing impairment) might need a different model for these variables. Judging samples in COPAS for these criteria would be a good start to attain these models.

Opportunities of the automatic analysis and the DIA tool are manifold. From the technical point of view, the DIA tool provides us with the means to collect more data, which allows us to build more robust and precise models. Of course, one has to take into account that those new data will be labeled by speech therapists whose labeling process may differ in many ways from the one that led to the COPAS labels from which the models were trained. One way to overcome this possible mismatch is to consider the difference between the score obtained by the new labeler and the score obtained by the current objective intelligibility model as a confidence measure. If this difference is below a threshold value, the perceptual score of the new labeler can be accepted and used as new training data for a stronger model.

Apart from the logical expansion of the DIA tool toward other pathologies, a first direction could be to develop a robust diagnosing system that composes a full speaker profile. From such a profile one could then retrieve objective information about the progress of a certain patient in the course of a therapy. In order to create robust models, a large dataset of well-annotated examples is necessary. COPAS is a very good start since it already contains speech of many pathological speakers. However, it is only judged by one human rater. This has two important consequences: (a) inter-rater agreement cannot be determined, so that there is no reference correlation to pursue, and (b) modeling one rater is dangerous since this means that the model will be subject to "noise" created by this rater. Since every human rater makes errors, it would be beneficial to have at least 5 ratings for every sample. In that case, the average rating would be sufficiently reliable. In order to investigate more sophisticated prediction models for monitoring a patient over time, more longitudinal recordings of a speaker, rated by several listeners for several perceptually measurable variables (intelligibility, articulation, phonation etc.) are needed.

Another direction of expanding the DIA tool could be to develop more universal models which can be used in many languages without needing much perceptually scored pathological speech samples from that language.

From a therapeutical point of view, an interesting challenge is to create a tool that can be used to steer the therapy. After a couple of general tests, such a tool could detect the problematic phonemes or articulatory dimensions and use this information to suggest exercises focusing on improving those dimensions or phonemes. An intersting investigation would then be to check whether these recommended exercises lead to positive results. Likewise, the DIA tool could also be useful as a tool for non-natives to learn Flemish or to correct for dialects.

In this respect, automatic visualizations of a patient's progress and in general in terms of his/her problematic phonological or phonetic dimensions could make the DIA tool both more appealing and more informative. Displaying properties of e.g. the vowel trapezium or, more general, situating a patient's realization of important articulatory dimensions against the "normal" range, could be very useful for the speech therapist. Information for these dimensions could come from a complete tool incorporating running speech and phoneme intelligibility tests as well as other tests such as the diadochokinetic rate, sustained vowel etc, all helping the speech therapist in gathering the global picture of a patient's possibilities.

Of course, the current DIA tool is just a first small step in this direction, but everything starts with a small step.

A
Phonetic alphabets

This appendix gives an overview of all Flemish and English phones (except for the closures and glottis closure) and their symbols according to several phonetic alphabets. Table A.1 presents all Flemish vowels and Table A.2 presents all Flemish consonants. Every phoneme is illustrated with a Flemish example. Table A.3 presents all English vowels and Table A.4 presents all English consonants. Every phoneme is illustrated with an English example.

IPA	YAPA	CGN	SAMPA	example
i	i	i	i	liep
ɪ	I	I	I	lip
e	e	e	e:	leeg
ɛ	E	E	E	leg
ɑ	A	A	A	lat
y	y	y	y	buut
ʏ	Y	Y	Y	put
ø	&	2	2:	deuk
ə	@	@	@	de
u	u	u	u	boek
o	o	o	o:	boom
ɔ	O	O	O	bom
a	a	a	a:	laat
ɛ̃	E~	E~	/	vaccin
ɔ̃	O~	O~	/	congé
ɑ̃	A~	A~	/	croissant
ɛː	E:	E:	E:	scène
ɔː	O:	O:	O:	zone (frans)
œː	@:	9:	9:	freule (frans)
ɛĭ	E:j/Eˆ	E+	Ei	wijs (diftong)
ɔŭ	O:w/Oˆ	A+	Au	koud (diftong)
œŭ	@:9/@ˆ	Y+	9y	huis (diftong)

Table A.1: Flemish vowels according to several phonetic alphabets.

IPA	YAPA	CGN	SAMPA	example
p	p	p	p	put
b	b	b	b	bad
t	t	t	t	tak
d	d	d	d	dak
k	k	k	k	kat
ɡ	g	g	g	goal
f	f	f	f	fop
v	v	v	v	vod
s	s	s	s	sap
z	z	z	z	zak
ʃ	S	S	S	sjaal
ʒ	Z	Z	Z	ravage
x	x	x	x	licht
ɣ	G	G	G	geen
h	h	h	h	heel
m	m	m	m	maan
n	n	n	n	nam
ŋ	N	N	N	lang
ɲ	Jj	J	/	oranje
l	l	l	l	loop
r	r	r	r	rook
j	j	j	j	ja
w/ʊ	w	w	w	weer

Table A.2: Flemish consonants according to several phonetic alphabets.

ARPABET	IPA	SAMPA	example
aa	ɒ,ɑ	Q,A	lock
ae	æ	{	bat
ah	ʌ	V	but
ao	ɔː	O	bought
aw	aʊ	aU	down
ax	ə	@	the
ax-h	əʰ		suspect
ax-r	ɚ		er butter
ay	aɪ	aI	buy
eh	e	E	bet
el	ᵊl		battle
em	ᵊm		bottom
en	ᵊn		button
eng	ᵊŋ		Washington
er	ɜː	3‘	bird
ey	eɪ	eI	bait
ih	ɪ	I	bits
ix			roses
iy	i	i	beat
ow	əʊ	@U	boat
oy	ɔɪ	OI	boy
uh	ʊ	U	book
uw	u	u	boot
ux			too
y	j	j	you

Table A.3: English vowels according to several phonetic alphabets.

ARPABET	IPA	SAMPA	example
b	b	b	bet
ch	ʧ	tS	church
d	d	d	door
dh	ð	D	that
dx			batter
f	f	f	fat
g	ɡ	g	get
hh	h	h	hat
hv	ɦ		(voiced h)
jh	ʤ	dZ	judge
k	k	k	kit
l	l	l	let
m	m	m	met
n	n	n	now
ng	ŋ	N	sing
nx			winter
p	p	p	pet
r	ɻ	r	rent
s	s	s	sat
sh	ʃ	S	shut
t	t	t	ten
th	θ	T	thing
v	v	v	vat
w	w	w	with
z	z	z	zoo
zh	ʒ	Z	pleasure

Table A.4: English consonants according to several phonetic alphabets.

B
Test material in COPAS

This appendix gives an overview of the texts of test material in COPAS I used in this dissertation.

B.1 DIA

There are 625 possible DIA tests. One set of 50 words can be found here.

List A (testing initial consonant): wop; piep; guis; leek; joen; kom; sar; baam; vil; roos; on; meur; vee; faf; toes; doon; nuk; hoef; zout
List B (testing final consonant): geep; diet; zoef; daam; jong; peeg; zaag; paai; tik; van; bool; lieuw; roor; toe; ries
List C (testing medial vowel): guil; zaat; dit; wiek; wan; hun; noet; vos; muul; woul; soos; teek; eut; rijd; man; del

B.2 Papa en Marloes

Papa en Marloes staan op het station. Ze wachten op de trein. Eerst hebben ze een kaartje gekocht. Er stond een hele lange rij, dus dat duurde wel even. Nu wachten ze tot de trein eraan komt. Het is al vijf over drie, dus het duurt nog vier minuten. Er staan nog veel meer mensen te wachten. Marloes kijkt naar links, in de verte ziet ze de trein al aankomen.

B.3 Sentences

Wil je liever de thee of de borrel?
Na nieuwjaar was hij weeral hier.

Bibliography

[1] K. M. Yorkston, P. A. Dowden, and D. R. Beukelman, "Intelligibility measurement as a tool in the clinical management of dysarhric speakers," in *Intelligibility in speech disorders: theory, measurement, and management*, ser. Studies in speech pathology and clinical linguistics, R. D. Kent, Ed. John Benjamins B. V., 1992.

[2] G. Van Nuffelen, "Speech Intelligibility in Dysarthria - Assessment and Treatment," Ph.D. dissertation, Universiteit Antwerpen, 2009.

[3] A. Maier, T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster, and E. Nöth, "PEAKS - A system for the automatic evaluation of voice and speech disorders," *Speech Communication*, vol. 51, pp. 425–437, 2009.

[4] J. N. Carmichael, "Introducing Objective Acoustic Metrics for the Frenchay Dysarthria Assessment Procedure," Ph.D. dissertation, University of Sheffield, 2007.

[5] http://www.esat.kuleuven.be/psi/spraak/projects/SPACE/.

[6] M. S. De Bodt, C. Guns, and G. Van Nuffelen, *NSVO: Nederlandstalig Spraakverstaanbaarheidsonderzoek*. Herentals: Vlaamse Vereniging voor Logopedisten, 2006.

[7] International Phonetic Association (IPA), *Handbook of the International Phonetic Association*. Cambridge University Press, 1999.

[8] http://www.phon.ucl.ac.uk/home/sampa/.

[9] I. Schuurman, M. Schouppe, H. Hoekstra, and T. van der Wouden, "CGN, an Annotated Corpus of Spoken Dutch," in *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03)*, 2003.

[10] D. H. Klatt, "Review of the ARPA Speech Understanding Project," *Journal of the Acoustical Society of America*, vol. 62, no. 6, pp. 1345–1366, 1977.

[11] F. Stouten, "Feature Extraction and Event Detection for Automatic Speech Recognition ," Ph.D. dissertation, Universiteit Gent, 2008.

[12] G. Van Nuffelen, C. Middag, M. S. De Bodt, and J. P. Martens, "Speech technology based assessment of phoneme intelligibility in dysarthria," *International Journal of Language and Communication Disorders*, vol. 44, no. 5, pp. 716–730, 2009.

[13] C. Middag, J. P. Martens, G. Van Nuffelen, and M. S. De Bodt, "Automated Intelligibility Assessment of Pathological Speech Using Phonological Features," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, p. 9, 2009.

[14] C. Middag, R. P. Clapham, R. van Son, and J. P. Martens, "Robust automatic intelligibility assessment techniques evaluated on speakers treated for head and neck cancer," *Computer Speech and Language*, submitted.

[15] G. Van Nuffelen, C. Middag, J. P. Martens, and M. S. De Bodt, "Speech technology based assessment of dysarthric speech : preliminary results," in *International Association of Logopedics and Phoniatrics, 27th World congress, Proceedings*. International Association of Logopedics and Phoniatrics (IALP), 2007, p. 5.

[16] C. Middag, G. Van Nuffelen, J. P. Martens, and M. S. De Bodt, "Objective intelligibility assessment of pathological speakers," in *Proceedings of the International Conference on Spoken Language Processing, Brisbane, Australia*, 2008, pp. 1745–1748.

[17] C. Middag, J. P. Martens, G. Van Nuffelen, , and M. S. De Bodt, "DIA: a tool for objective intelligibility assessment of pathological speech," in *Proceedings of the 6th International Workshop for Models and Analysis of Vocal Emissions for Biomedical Applications*, 2009, p. 4.

[18] C. Middag, Y. Saeys, and J. P. Martens, "Towards an ASR-Free Objective Analysis of Pathological Speech," in *Proceedings of the International Conference on Spoken Language Processing, Tokio, Japan*, 2010, pp. 294–297.

[19] C. Middag, T. Bocklet, J. P. Martens, and E. Nöth, "Combining phonological and acoustic ASR-free features for pathological speech intelligibility assessment," in *Proceedings of the International Conference on Spoken Language Processing, Florence, Italy*, 2011, pp. 3005–3008.

[20] R. Dharmaperwira-Prins, *Dysartrie en verbale apraxie: beschrijving, onderzoek, behandeling*. Swets & Zeitlinger Publishers, 1996.

[21] http://www.sylvius.com/.

[22] D. B. Freed, *Motor speech disorders: diagnosis and treatment*. Cengage Learning, 2000, vol. 1.

[23] L. Morris and S. Afifi, *Tracheostomies: The Complete Guide*. Springer, 2010.

[24] B. H. Story, "An overview of the physiology, physics, and modeling of the sound source for vowels," *Acoustical Science and Technology*, vol. 23, no. 4, pp. 195–206, 2002.

[25] H. K. Schutte, "Fysiologie van de stemgeving," in *Handboek stem-spraak-taalpathologie*, H. F. M. Peters, R. Bastiaanse, J. van Borsel, P. H. O. Dejonckere, K. Jansonius-Schultheiss, S. van der Meulen, and B. J. E. Mondelaers, Eds. Bohn Stafleu Van Loghum, 1999, ch. A3.1.1, pp. 1–37.

[26] http://www.hoofdhalskanker.info/.

[27] I. Titze, *Principles of voice production*. Englewood Cliffs, New Jersey: Prentice Hall, 1994.

[28] R. J. Baken and R. F. Orlikoff, *Clinical Measurement of Speech and Voice*. Singular Thomson Learning, 2000.

[29] A. C. M. Rietveld and V. J. van Heuven, *Algemene fonetiek*. Coutinho, 1997.

[30] M. Moerman and H. Vermeersch, "Maligne larynxpathologie - de behandeling," in *Handboek stem-spraak-taalpathologie*, H. F. M. Peters, R. Bastiaanse, J. van Borsel, P. H. O. Dejonckere, K. Jansonius-Schultheiss, S. van der Meulen, and B. J. E. Mondelaers, Eds. Bohn Stafleu Van Loghum, 2003, ch. B1.3.7.1, pp. 1–16.

[31] J. Bernier, Ed., *Head and Neck Cancer: Multimodality Management*. Springer, 2011.

[32] A. Olthoff, S. Mrugalla, R. Laskawi, M. Froehlich, I. Stuermer, E. Kruse, P. Ambrosch, and W. Steiner, "Assessment of Irregular Voices after Total and Laser Surgical Partial Laryngectomy," *Archives of Otolaryngology - Head & Neck Surgery*, vol. 129, no. 9, pp. 994–999, 2003.

[33] J. P. Dworkin, R. J. Meleca, M. A. Zacharek, R. J. Stachler, R. Pasha, G. G. Abkarian, R. A. Culatta, and J. R. Jacobs, "Voice and deglutition functions after the supracricoid and total laryngectomy procedures for advanced stage laryngeal carcinoma," *Otolaryngology - Head and Neck Surgery*, vol. 129, no. 4, pp. 311–320, 2003.

[34] G. Torrejano and I. Guimaraes, "Voice Quality After Supracricoid Laryngectomy and Total Laryngectomy With Insertion of Voice Prosthesis," *Journal of Voice*, vol. 23, no. 2, pp. 240–246, 2009.

[35] Y. K. So, Y. S. Yun, C. H. Baek, H. S. Jeong, and Y. I. Son, "Speech outcome of supracricoid partial laryngectomy: Comparison with total laryngectomy and anatomic considerations," *Otolaryngology - Head and Neck Surgery*, vol. 141, no. 6, pp. 770–775, 2009.

[36] J. Pfuetzenreiter, A. D. Dedivitis, D. S. Queija, N. P. Bohn, and A. P. B. Barros, "The Relationship Between the Glottic Configuration After Frontolateral Laryngectomy and the Acoustic Voice Analysis," *Journal of Voice*, vol. 24, 2009.

[37] A. Singh, R. Kazi, J. De Cordova, C. M. Nutting, P. Clarke, K. J. Harrington, and P. RhysEvans, "Multidimensional Assessment of Voice After Vertical Partial Laryngectomy: A Comparison with Normal and Total Laryngectomy Voice," *Journal of Voice*, vol. 22, no. 6, pp. 740–745, 2008.

[38] H. A. Van Wijngaarden and H. P. J. Moes, "Oesofagale stem," in *Handboek stem-spraak-taalpathologie*, H. F. M. Peters, R. Bastiaanse, J. van Borsel, P. H. O. Dejonckere, K. Jansonius-Schultheiss, S. van der Meulen, and B. J. E. Mondelaers, Eds. Bohn Stafleu Van Loghum, 2003, ch. B1.3.7.2a, pp. 1–24.

[39] M. S. Weiss, G. H. Yeni-Komshian, and J. M. Heinz, "Acoustical and perceptual characteristics of speech produced with an electronic artificial larynx," *Journal of the Acoustical Society of America*, vol. 65, no. 5, pp. 1298–1308, 1979.

[40] H. Liu and M. L. Ng, "Electrolarynx in voice rehabilitation," *Auris Nasus Larynx*, vol. 34, pp. 327–332, 2007.

[41] D. H. Brown, F. J. M., Hilgers, J. C. Irish, and A. J. M. Balm, "Postlaryngectomy Voice Rehabilitation: State of the Art at the Millennium," *World Journal of Surgery*, vol. 27, no. 7, pp. 824–31, 2003.

[42] F. J. M. Hilgers and C. J. Van As, "Spraakrevalidatie na totale laryngectomie door middel van stemprothesem," in *Handboek stem-spraak-taalpathologie*, H. F. M. Peters, R. Bastiaanse, J. van Borsel, P. H. O. Dejonckere, K. Jansonius-Schultheiss, S. van der Meulen, and B. J. E. Mondelaers, Eds. Bohn Stafleu Van Loghum, 2002, ch. B1.3.7.2b, pp. 1–30.

[43] T. Most, Y. Tobin, and R. C. Mimran, "Acoustic and perceptual characteristics of esophageal and tracheoesophageal speech production," *Journal of Communication Disorders*, vol. 33, no. 2, pp. 165–181, 2000.

[44] W. M. Diedrich and K. A. Youngstrom, *Alaryngeal Speech*. Thomas, 1966.

[45] J. P. Searl and P. M. Evitts, "Velopharyngeal aerodynamics of /m/ and /p/ in tracheoesophageal speech," *Journal of Voice*, vol. 18, no. 4, pp. 557–566, 2004.

[46] H. Liu, M. L. Ng, M. Wan, S. Wang, and Y. Zhang, "Effects of place of articulation and aspiration on voice onset time in Mandarin esophageal speech," *Folia Phoniatrica et Logopaedica*, vol. 59, no. 3, pp. 147–154, 2007.

[47] H. Hirose, "Voicing Distinction in Esophageal Speech," *Acta Oto-Laryngologica*, vol. Supplement 524, pp. 56–63, 1996.

[48] P. J. Watson and R. S. Schlauch, "Fundamental Frequency Variation With an Electrolarynx Improves Speech Understanding: A Case Study," *American Journal of Speech-Language Pathology*, vol. 18, no. 2, pp. 162–167, 2009.

[49] H. S. Choi, Y. J. Park, S. M. Lee, and K. M. Kim, "Functional Characteristics of a New Electrolarynx "Evada" having a Force Sensing Resistor Sensor," *Journal of Voice*, vol. 15, no. 4, pp. 592–599, 2001.

[50] S. E. Williams and J. B.Watson, "Speaking proficiency variations according to method of alaryngeal voicing," *The Laryngoscope*, vol. 97, no. 6, pp. 737–739, 1987.

[51] J. P. Searl, M. A. Carpenter, and C. L. Banta, "Intelligibility of stops and fricatives in tracheoesophageal speech," *Journal of Communication Disorders*, vol. 34, no. 4, pp. 305–321, 2001.

[52] F. J. M. Hilgers, A. H. Ackerstaff, and C. J. Van As, "Tracheoesophageal puncture: prosthetic voice management," *Current Opinion in Otolaryngology & Head and Neck Surgery*, vol. 7, no. 3, pp. 112 –118, 1999.

[53] T. Haderlein, "Automatic Evaluation of Tracheoesophageal Substitute Voices," Ph.D. dissertation, Universität Erlangen-Nürnberg, 2007.

[54] H. F. Mahieu, *Voice and speech rehabilitation following laryngectomy*. Universiteit Groningen, 1988.

[55] P. Jongmans, F. J. M. Hilgers, L. C. Pols, and C. J. van As-Brooks, "The intelligibility of tracheoesophageal speech, with an emphasis on the voiced-voiceless distinction," *Logopedics Phoniatrics Vocology*, vol. 31, no. 4, pp. 172–181, 2006.

[56] H. A. Van Wijngaarden, "Schisis en veluminsufficiëntie - anatomische aspecten," in *Handboek stem-spraak-taalpathologie*, H. F. M. Peters, R. Bastiaanse, J. van Borsel, P. H. O. Dejonckere, K. Jansonius-Schultheiss, S. van der Meulen, and B. J. E. Mondelaers, Eds. Bohn Stafleu Van Loghum, 2004, ch. B3.1.2b, pp. 1–26.

[57] K. Jansonius-Schultheiss and E. Konst, "Spraakverwerving van baby's en peuters met een schisis," in *Handboek stem-spraak-taalpathologie*, H. F. M. Peters, R. Bastiaanse, J. van Borsel, P. H. O. Dejonckere, K. Jansonius-Schultheiss, S. van der Meulen, and B. J. E. Mondelaers, Eds. Bohn Stafleu Van Loghum, 2003, ch. B3.1.2.1, pp. 1–19.

[58] R. Wyatt and D. Sell and J. Russell and A. Harding and K. Harland and L. Albery, "Cleft palate speech dissected: a review of current knowledge and analysis," *British Journal of Plastic Surgery*, vol. 49, no. 3, pp. 143 – 149, 1996.

[59] J. E. Trost, "Articulatory Additions to the Classical Description of the Speech of Persons with Cleft Palate," *Journal of Cleft Palate*, vol. 18, pp. 193–198, 1981.

[60] K. M. Van Lierde and S. Monstrey and K. Bonte and P. Van Cauwenberge and B. Vinck, "The long-term speech outcome in Flemish young adults after two different types of palatoplasty," *International Journal of Pediatric Otorhinolaryngology*, vol. 68, no. 7, pp. 865 – 875, 2004.

[61] B. A. Stach, *Clinical Audiology: an Introduction*. Cengage Learning, 2010.

[62] M. J. Osberger, "Intelligibility in the Hearing Impaired: Research and Clinical Implications," in *Intelligibility in speech disorders: theory, measurement, and management*, ser. Studies in speech pathology and clinical linguistics, R. D. Kent, Ed. John Benjamins B. V., 1992.

[63] J. M. Lenden. and P. Flipsen Jr, "Prosody and voice characteristics of children with cochlear implants," *Journal of Communication Disorders*, 2006.

[64] R. D. Kent, Ed., *Intelligibility in Speech Disorders: Theory, Measurement, and Management*, ser. Studies in speech pathology and clinical linguistics. John Benjamins B.V., 1992.

[65] R. Likert, "A technique for the measurement of attitudes," *Archives of Psychology*, vol. 22, no. 140, pp. 1–55, 1932.

[66] G. Van Nuffelen and M. S. De Bodt, "Report on the study of different tests developed for the assessment of articulation and speech intelligibility - Proposal for a test able to take advantage of the opportunities offered by the use of an articulatory speech analyzer and/or a phonetic speech recognizer," Universiteit Antwerpen, Tech. Rep., 2006.

[67] P. R. Shrout and J. L. Fleiss, "Intraclass Correlations: Uses in Assessing Rater Reliability," *Psychological Bulletin*, vol. 86, pp. 420–428, 1979.

[68] G. Van Nuffelen, M. S. De Bodt, F. Wuyts, and P. Van de Heyning, "Reliability and Clinical Relevance of a Segmental Analysis based on an Intelligibility Assessment," *Folia Phoniatrica et Logopaedica*, vol. 60, pp. 264–268, 2008.

[69] L. Ferrier, H. Shane, H. Ballard, T. Carpenter, and A. Benoit, "Dysarthric Speakers' Intelligibility and Speech Characteristics in Relation to Computer Speech Recognition," *Journal of Augmentative and Alternative Communication*, vol. 11, pp. 165–74, 1995.

[70] P. Vijayalakshmi, R. Reddy, and D. O'Shaughnessy, "Assessment of Articulatory Sub-systems of Dysarthric Speech Using an Isolated-style Phoneme Recognition System," in *Proceedings of the International Conference on Spoken Language Processing*, 2006, pp. 981–984.

[71] P. Vijayalakshmi, T. Nagarajan, and M. R. Reddy, "Assessment of Articulatory and Velopharyngeal Sub-systems of Dysarthric Speech," *International Journal of Biomedical Soft Computing and Human Sciences, special issue on Biosensors: Data acquisition, Processing and Control*, vol. 14, no. 2, pp. 87–94, 2009.

[72] J. P. Hosom, L. Shriberg, and J. Green, "Diagnostic Assessment of Childhood Apraxia of Speech Using Automatic Speech Recognition (ASR) Methods," in *Journal ofMedical Speech Language Pathology*, vol. 12, no. 4, 2004, pp. 167–171.

[73] L. Gu, J. G. Harris, R. Shrivastav, and C. Sapienza, "Disordered Speech Assessment Using Automatic Methods based on Quantitative Measures," *EURASIP Journal on Applied Signal Processing*, vol. 9, pp. 1400–1409, 2005.

[74] H. Su, C. Wu, and P. J. Tsai, "Automatic assessment of articulation disorders using confident unit-based model adaptation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2008, pp. 4513–4516.

[75] T. H. Falk, W. Y. Chan, and F. Shein, "Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility," *Speech Communication*, vol. 54, no. 5, pp. 622–631, 2011.

[76] P. Enderby, "Frenchay dysarthria assessment," *International Journal of Language & Communication Disorders*, vol. 15, no. 3, pp. 165–173, 1980.

[77] A. V. Fox, *PLAKSS: psycholinguistische Analyse kindlicher Sprechstörungen*. Harcourt Test Services, 2005.

[78] G. Stemmer, C. Hacker, S. Steidl, and E. Nöth, "Acoustic Normalization of Children's Speech," in *Proceedings of the European Conference on Speech Communication and Technology*, Eurospeech, Ed., vol. 2, 2003, pp. 1313–1316.

[79] A. Maier, "Speech of Children with Cleft Lip and Palate: Automatic Assessment," Ph.D. dissertation, Universität Erlangen-Nürnberg, 2009.

[80] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, ser. Springer Series in Statistics. Springer New York Inc., 2001.

[81] I. T. Jolliffe, *Principal Component Analysis*, ser. Springer Series in Statistics. Springer-Verlag, 2002.

[82] J. H. Holland, *Adaptation in Natural and Artificial Systems: an Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. University of Michigan Press, 1975.

[83] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[84] L. Breiman, "Bagging Predictors," in *Machine Learning*, vol. 24, 1996, pp. 123–140.

[85] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.

[86] M. Varewyck and J. P. Martens, "A practical approach to model selection for support vector machines with a gaussian kernel," *IEEE transactions on systems man and cybernetics part B - cybernetics*, 2011. [Online]. Available: http://dx.doi.org/10.1109/TSMCB.2010.2053026

[87] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," in *Proceedings of IEEE*, vol. 77, 1989, pp. 257–286.

[88] J. L. Gauvain and C. H. Lee, "Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

[89] H. Daumé III, "Frustratingly easy domain adaptation," in *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, Prague, Czech Republic, 2007.

[90] J. P. Martens, *Spraakverwerking*. Ghent University, 2006.

[91] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-28, no. 4, pp. 357–366, 1980.

[92] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-Based State Tying for High Accuracy Acoustic Modelling," in *Proceedings of the ARPA Human Language Technology Workshop, Plainsboro*, 1994.

[93] K. Demuynck, D. Van Compernolle, C. Van Hove, and J. P. Martens, *Een Corpus gesproken Nederlands voor spraaktechnologisch Onderzoek. Final Report of CoGeN Project*. ELIS UGent, Gent, 1997.

[94] J. Lernout and P. Hauspie, "Automatic Speech Recognition ASR1500/ASR1600 - Software Development Kit for Windows 95/98/NT," Lernout & Hauspie Speech Products, Tech. Rep., 1999.

[95] G. Van Nuffelen, M. S. De Bodt, C. Middag, and J. P. Martens, "Dutch corpus of pathological and normal speech (copas)," Antwerp University Hospital and Ghent University, Tech. Rep., 2009.

[96] J. Van De Weijer and I. Slis, "Nasaliteitsmeting met de nasometer," *Logopedie en Foniatrie*, vol. 63, pp. 97–101, 1991.

[97] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (Version 4.4.34) [Computer program]. Retrieved October 19, 2006 from http://www.praat.org/ ," 2006.

[98] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2011, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org

[99] J. P. Martens, D. Binnenpoorte, K. Demuynck, R. Van Parys, T. Laureys, W. Goedertier, and J. Duchateau, "Word Segmentation in the Spoken Dutch Corpus," in *Proceedings of the 3th international conference on Language Resources and Evaluation (LREC)*, vol. V, Las Palmas, Canary Islands, Spain, May 2002, pp. 1432–1437.

[100] K. Demuynck and T. Laureys and P. Wambacq and D. Van Compernolle, "Automatic Phonemic Labeling and Segmentation of Spoken Dutch," in *Proceedings of the 4th international conference on Language Resources and Evaluation (LREC)*, vol. I, Lisbon, Portugal, May 2004, pp. 61–64.

[101] D. M. Binnenpoorte, "Phonetic transcriptions of large speech corpora," Ph.D. dissertation, Universiteit Nijmegen, 2006.

[102] J. Carmichael and P. Green, "Revisiting Dysarthria Assessment Intelligibility Metrics," in *8th International Conference on Spoken Language Processing (ICSLP) October 4-8, Korea*, 2004, pp. 742–745.

[103] F. Stouten and J. P. Martens, "On the Use of Phonological Features for Pronunciation Scoring," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2006, pp. 329–332.

[104] K. Demuynck, "Extracting, modelling and combining information in speech recognition," Ph.D. dissertation, K.U.Leuven, ESAT, 2001.

[105] D. Van Compernolle, "Noise Adaptation in a Hidden Markov Model Speech Recognition System," *Computer Speech and Language*, vol. 3, no. 2, pp. 151–168, 1989.

[106] K. Demuynck and J. Duchateau and D. Van Compernolle, "Optimal Feature Sub-space Selection based on Discriminant Analysis," in *Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECH)*, vol. III, Budapest, Hungary, Sep. 1999, pp. 1311–1314.

[107] K. Demuynck and J. Duchateau and D. Van Compernolle and P. Wambacq, "Improved Feature Decorrelation for HMM-based Speech Recognition," in *Proceedings of the International Conference on Spoken Language Processing*, vol. VII, Sydney, Australia, Dec. 1998, pp. 2907–2910.

[108] W. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA Speech Recognition Research Database: Specifications and Status," in *DARPA Workshop on Speech Recognition*, 1986, pp. 93–99.

[109] K. Demuynck and J. Roelens and D. Van Compernolle and P. Wambacq, "SPRAAK: An Open Source SPeech Recognition and Automatic Annotation Kit," in *Proceedings of the International Conference on Spoken Language Processing*, Brisbane, Australia, Sep. 2008, p. 495.

[110] K. Riedhammer, G. Stemmer, T. Haderlein, M. Schuster, F. Rosanowski, E. Nöth, and A. Maier, "Towards Robust Automatic Evaluation of Pathologic Telephone Speech," in *Proceedings of the Workshop on Automatic Speech Recognition and Understanding*, IEEE, Ed., vol. 1, 2007, pp. 717–722.

[111] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*. CRC Press, 2004.

[112] M. Galassi, J. Davies, J. Theiler, B. Gough, G. Jungman, P. Alken, M. Booth, and F. Rossi, *GNU Scientific Library Reference Manual* , 3rd ed., B. Gough, Ed. Network Theory Ltd., 2009.

[113] C. Chang and C. J. Lin, *LIBSVM: a Library for Support Vector Machines*, 2001. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm

[114] K. Tjaden, "Segmental Articulation in Motor Speech Disorders," in *Motor Speech Disorders*, G. Weismer, Ed. San Diego, CA: Plural, 2007, pp. 151–186.

[115] A. T. Neel, "Vowel Space Characteristics and Vowel Identification Accuracy," *Journal of Speech, Language, and Hearing Research*, vol. 51, no. 3, p. 574, 2008.

[116] B. M. Ansel and R. D. Kent, "Acoustic-Phonetic Contrasts and Intelligibility in the Dysarthria Associated With Mixed Cerebral Palsy," *Journal of Speech and Hearing Research*, vol. 35, no. 2, pp. 296–308, 1992.

[117] R. D. Kent, J. F. Kent, G. Weismer, R. L. Sufit, J. C. Rosenbek, R. E. Martin, and B. R. Brooks, "Impairment of Speech Intelligibility in Men with Amyotrophic Lateral Sclerosis," *Journal of Speech and Hearing Disorders*, vol. 55, no. 4, pp. 721–728, 1990.

[118] M. J. Osberger and N. S. McGarr, "Speech Production Characteristics of the Hearing Impaired," in *Speech and Language: Advances in Basic Research and Practice*, ser. Speech and Language, N. J. Lass, Ed. Academic Press, 1982, vol. 8, pp. 221–283.

[119] H. Liu, F. Tsao, and P. K. Kuhl, "The effect of reduced vowel working space on speech intelligibility in Mandarin-speaking young adults with cerebral palsy," *Journal of Acoustic Society of America*, vol. 117, no. 6, pp. 3879–89, 2005.

[120] H. Chen and K. N. Stevens, "An acoustical study of the fricative /s/ in the speech of individuals with dysarthria." *Journal of Speech, Language, and Hearing Research*, vol. 44, no. 6, pp. 1300–14, 2001.

[121] R. D. Kent, G. Weismer, J. F. Kent, and J. C. Rosenbek, "Toward Phonetic Intelligibility Testing in Dysarthria," *Journal of Speech and Hearing Disorders*, vol. 54, pp. 482–499, 1989.

[122] L. Uvin, "De correlatie tussen objectieve en perceptuele spraakverstaanbaarheidsscores bij pathologische spraak," Master's thesis, Universiteit Gent, Belgium, 2009.

[123] S.G. Fletcher and D.A. Daly, "Nasalance in utterances of hearing-impaired speakers," *Journal of Communication Disorders*, vol. 9, no. 1, pp. 63 – 73, 1976.

[124] S. B. Leder and J. B. Spitzer, "A Perceptual Evaluation of the Speech of Adventitiously Deaf Adult Males," *Ear and hearing*, vol. 11, no. 3, pp. 169–75, 1990.

[125] J. Duchateau, K. Demuynck, and H. Van Hamme, "Evaluation of phone lattice based speech decoding," in *Proceedings of the European Conference on Speech Communication and Technology, Brighton, U.K.*, 2009, pp. 1179–1182.

[126] T. Bocklet, T. Haderlein, F. Hönig, F. Rosanowski, and E. Nöth, "Evaluation and Assessment of Speech Intelligibility on Pathologic Voices based upon Acoustic Speaker Models," in *Proceedings of the 3rd Advanced Voice Function Assessment International Workshop*, 2009, pp. 89–92.

[127] C. Spearman, "The proof and measurement of association between two things. By C. Spearman, 1904." *The American journal of psychology*, vol. 100, no. 3-4, pp. 441–471, 1987.

[128] I. H. Witten, E. Frank, L. Trigg, M. Hall, G. Holmes, and S. J. Cunningham, "Weka: Practical Machine Learning Tools and Techniques with Java Implementations," in *Proceedings of the ICONIP/ANZIIS/ANNES'99 Workshop on Emerging Knowledge Engineering and Connectionist-Based Information Systems*, 1999, pp. 192–196.

[129] L. van der Molen, M. A. van Rossum, I. Jacobi, R. J. J. H. van Son, L. E. Smeele, C. R. N. Rasch, and F. J. M. Hilgers, "Pre- and posttreatment voice and speech outcomes in patients with advanced head and neck cancer treated with chemoradiotherapy: Expert listeners' and patient's perception," *Journal of Voice*, vol. online, January 3 2012.

[130] R. P. Clapham, L. van der Molen, R. J. J. H. van Son, M. W. M. van den Brekel, and F. J. M. Hilgers, "NKI-CCRT Corpus: Speech Intelligibility Before and After Advanced Head and Neck Cancer Treated with Concomitant Chemoradiotherapy," in *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC)*, 2012, p. In Press.

[131] D. Van Compernolle, J. Smolders, P. Jaspers, and T. Hellemans, "Speaker Clustering for Dialectic Robustness in Speaker Independent Speech Recognition," in *Proceedings of the European Conference on Speech Communication and Technology,Genova, Italy*, 1991, pp. 723–726.

[132] J. Nerbonne, W. Heeringa, E. van den Hout, P. van der Kooi, S. Otten, W. van de Vis, and Alfa-informatica Bcn, "Phonetic Distance between Dutch Dialects," in *Proceedings of the Computational Linguistics in the Netherlands meeting, Antwerp*. Available, 1995, pp. 185–202.

[133] M. J. de Bruijn, L. ten Bosch, D. J. Kuik, H. Quené, J. A. Langendijk, and C. R. L. I. M. Verdonck-de Leeuw, "Objective Acoustic-Phonetic Speech Analysis in Patients Treated for Oral or Oropharyngeal Cancer," *Folia Phoniatrica et Logopaedica*, vol. 61, no. 3, pp. 180–187, 2009.

[134] M. S. De Bodt, H. M. E. Hernández-Díaz, and P. H. Van De Heyning, "Intelligibility as a Linear Combination of Dimensions in Dysarthric Speech," *Journal of Communication Disorders*, vol. 35, no. 3, pp. 283–92, 2002.

[135] I. Jacobi, "On Variation and Change in Dipthongs and Long Vowels of Spoken Dutch," Ph.D. dissertation, University of Amsterdam, 2009.

[136] M. J. Munro, "Foreign accent and speech intelligibility," in *Phonology and second language acquisition*, ser. Studies in Bilingualism, E. Hansens, G. Jette, and M. L. Zampini, Eds. John Benjamins, 2008, vol. 36, pp. 193–218.

[137] S. van Wijngaarden, "Intelligibility of native and non-native Dutch speech," *Speech Communication*, vol. 35, no. 1-2, pp. 103–113, 2001.

[138] L. van Immerseel and J. P. Martens, "AMPEX Disordered Voice Analyzer [computer program]," Digital Speech and Signal Processing research group Ghent University, Belgium, version date July 2008 -. [Online]. Available: http://speech.elis.ugent.be/

[139] M. Moerman, G. Pieters, J. P. Martens, M. J. Van der Borgt, and P. Dejonckere, "Objective evaluation of the quality of substitution voices," *European Archives of Oto-Rhino-Laryngology*, vol. 261, no. 10, pp. 541–547, 11 2004.

[140] L. Van Immerseel and J. P. Martens, "Pitch and Voiced/Unvoiced Determination with an Auditory Model," *Journal of the Acoustical Society of America*, vol. 91, no. 6, pp. 3511–3526, 1993.

[141] L. A. Newman, K. T. R. J. A. Logemann, A. W. Rademaker, C. L. Lazarus, A. Hamner, S. Tusant, and C. F. Huang, "Swallowing and speech ability after treatment for head and neck cancer with targeted intraarterial versus intravenous chemoradiation," *Head & Neck*, vol. 24, pp. 68–77, 2001.

[142] A. Neri, C. C. Cucchiarini, and H. Strik, "Selecting Segmental Errors in Non-native Dutch for Optimal Pronunciation Training," *International Review of Applied Linguistics*, vol. 44, pp. 357–404, 2006.

[143] I. Jacobi, L. van der Molen, H. Huiskens, M. A. van Rossum, and F. J. M. Hilgers, "Voice and Speech Outcomes of Chemoradiation for Advanced

Head and Neck Cancer: A Systematic Review," *European Archives of Oto-Rhino-Laryngology*, vol. online, 2010.

[144] D. Seynhaeve, "Ontwikkeling van een vaste set NSVO-testitems voor kinderen en volwassenen," Master's thesis, Universiteit Gent, Belgium, 2010.