Speech of Children with Cleft Lip and Palate: Automatic Assessment

Submitted to

Technische Fakultät der
Universität Erlangen-Nürnberg

in partial fulfillment of the requirements for
the degree of

DOKTOR-INGENIEUR

by

Andreas Maier

Erlangen — 2009

Deutscher Titel:

Sprache bei Kindern mit Lippen-Kiefer-Gaumenspalte: Automatische Bewertung

As dissertation accepted by
Technische Fakultät der
Universität Erlangen-Nürnberg

Date of submission:	4th September, 2008
Date of doctorate:	15th January, 2009
Dean:	Prof. Dr.-Ing. habil. J. Huber
Reviewer:	Prof. Dr.-Ing. habil. E. Nöth
	Prof. S. Furui (Ph. D.)

Acknowledgment

I'd like to thank my supervisor Prof. Dr.-Ing. Elmar Nöth who supported me throughout this work. It would not have been possible to complete this thesis without his support and his valuable comments on my algorithms and this manuscript.

During my stay in Japan, I had the opportunity to visit the lab of Prof. Sadaoki Furui (Ph. D.). Shortly after my visit, Prof. Furui agreed to review this thesis. This is a great honor for me.

PD Dr. med. Maria Schuster supported me with her broad knowledge and could answer all my questions concerning treatment and therapy of children with cleft lip and palate. Moreover, she did a lot of work in organization and preparation of the recordings and the perceptive assessment. Due to her efforts, it was possible to aquire this much data within our project. Furthermore, I thank Prof. Dr. med. Dr.-Ing. Ulrich Eysholdt and Prof. Dr. med. Frank Rosanowski for the opportunity to work at their division.

For the contribution to the data collection, I would like to express my gratitude to the Michael-Poeschke-Schule Erlangen, the Erna-Zink-Kindergarten Erlangen, the Sabel Schule Nuremberg, the Eichendorff-Schule Hannover, the Gutenberg Schule Karlsruhe, the Nebenius-Grundschule Karlsruhe, the Hans-Thoma Schule Karlsruhe, and the Dritte Schule Leipzig. I would like to thank the many helpers who contributed to this work by recording, segmenting, and transliterating the data:
Franziska Dames, Katja Becher, Christine Kiourkenidis-Martínez, Dipl.-Inf. Tobias Bocklet, Julian Exner, Erik Körner, Sebastian Korschofski, Meng Li, Eric Nöth, Alexander Reuß, Dr. med. dent. Florian Stelzle, Anika Ströbele, and Dr. med. dent. Beate Vogt.

I'd like to thank the speech therapists Andrea Schädel, Dorothee Großmann and Dr. phil. Ulrike Wohlleben for the experts' annotations of the data. Prof. Dr. Peter Kummer, PD Dr.-Ing. Jörg Lohscheller, Dr.-Ing. Martin Burger, and Dr.-Ing. Tobias Wurzbacher also contributed to this work by assessment of the intelligibility.

Moreover, I would like to thank Dipl.-Inf. Viktor Zeissler for writing the application for the six months of funding by the Johannes-und-Frieda-Marohn-Stiftung. Also, the DFG supported my work by the grant of the three-year project "Automatische Sprachanalyse von Kindern und Jugendlichen mit Lippen-Kiefer-Gaumenspalten".

For the helpful comments throughout my work, I would like to thank Dr. phil. Anton Batliner, Dipl.-Inf. Christian Hacker, Dr.-Ing. Tino Haderlein, Dipl.-Inf. Florian Hönig, PD Dr. med. Dr. med. dent. Emeka Nkenke, and with special thanks for the Sammon viewer code Dr.-Ing. Stefan Steidl and also thanks to Dipl.-Ing. Marcus Barkowski, Dipl.-Ing. Roland Bitto, and Dipl.-Ing. Björn Eskofier for the access to the Opticom 3SQM algorithms.

Finally, I want to thank my family and my wife Yoshimi who always kept patient when I was coming home late from work while writing this thesis.

Andreas Maier

Abstract

This work investigates the use of automatic speech processing techniques for the automatic assessment of children's speech disorders. The target group were children with cleft lip and palate (CLP). The speech processing techniques are applied to evaluate the children's speech intelligibility and their articulation. Another goal of this work is to visualize the kind and degree of the pathology in the children's speech. Tracking of the children's therapy progress is also within the reach of the system.

Cleft lip and palate is the most common orofacial alteration. Even after adequate surgery, speech and hearing is still affected. The articulation or speech disorders of the children consist of typical misarticulations such as backing of consonants and enhanced nasal air emission.

State-of-the-art evaluation of speech disorders is performed perceptively by human listeners. This method, however, is hampered by inter- and intra-individual differences. Therefore, an automatic evaluation is desirable.

We developed PEAKS — the **P**rogram for the **E**valuation of **A**ll **K**inds of **S**peech disorders. With PEAKS one can record and evaluate speech data via the Internet. It runs in any web browser and features security concepts such as secure transmission and user level access control.

The agreement of PEAKS with different human experts is measured with different correlation coefficients, Kappa, and Alpha. The evaluation procedures for intelligibility employ Support Vector Machines and Regression. Furthermore, dimensionality reduction techniques such as LDA, PCA, and Sammon mapping are used for the visualization and the feature reduction. As input for these algorithms typical speech processing features such as MFCCs as well as specialized feature sets for prosody, pronunciation, and hypernasalization are employed. Another approach of this work is to use a children's speech recognizer to model a naïve listener. If the recording conditions are kept constant, the speaker should be the only varying factor. Hence, the recognition rate should resemble the intelligibility of the speaker.

Collection of patient speech data was performed in Erlangen from 2002 until 2008. 312 children with CLP were recorded. Control groups were gathered in four major cities of Germany to cover several regions of dialect. 726 control data sets were acquired.

The experimental results showed that the automatic system yields a high and significant agreement to the human raters for global parameters such as intelligibility as well as single articulation disorders. The system is in the same range as the human raters. The intelligibility assessment was shown to be independent of the region of dialect. The visualization of the speech data also showed high agreement to perceptively rated criteria. Artifacts which were caused by the use of multiple microphones were removed.

Übersicht

Diese Arbeit untersucht die Verwendung von automatischen Sprachverarbeitungstechniken für die automatische Bewertung von Kindern mit Sprechstörungen. Die Zielgruppe waren Kinder mit Lippen-Kiefer-Gaumenspalte (LKG). Die Sprachverarbeitungstechniken werden verwendet, um die Verständlichkeit und die Artikulation der Kinder zu bewerten. Ein weiteres Ziel von dieser Arbeit ist die Visualisierung der Art und des Grades der Pathologie in der Kindersprache. Verlaufskontrolle der Therapie der Kinder ist ebenfalls innerhalb der Reichweite des Systems.

Lippen-Kiefer-Gaumenspalte ist die häufigste orofaziale Deformation. Auch nach ausreichender chirurgischer Behandlung sind Sprech- und Hörvermögen immer noch betroffen. Die Artikulations- oder Sprechstörungen der Kinder enthalten typische Fehlartikulationen wie Rückverlagerung von Konsonanten und nasale Luft-Emission bei Vokalen.

Bewertung von Sprechstörungen erfolgt durch menschliche Bewerter nach dem aktuellen Stand der Technik. Diese Methode unterliegt jedoch inter- und intraindividuellen Unterschieden. Daher wird eine automatische Auswertung ist wünschenswert.

Wir entwickelten PEAKS - Das Programm zur Evaluation und Analyse kindlicher Sprechstörungen. Mit PEAKS kann man Sprachdaten aufzeichnen und über das Internet auswerten. Es funktioniert in jedem Web-Browser und erfüllt Sicherheits-Konzepte wie die sichere Übertragung und Zugriffskontrolle auf Benutzer-Ebene.

Die Übereinstimmung von PEAKS mit verschiedenen menschlichen Experten wird mit verschiedenen Korrelations-Koeffizienten, Kappa, und Alpha bestimmt. Das Bewertungsverfahren für die Verständlichkeit benutzt Support Vector Maschinen und Regression. Darüber hinaus werden Techniken wie LDA, PCA, und Sammon Mapping verwendet und die Dimension für eine Visualisierung zu reduzieren. Als Eingabe für diese Algorithmen werden typische Sprachverarbeitungsmerkmale wie MFCCs sowie spezielle Merkmale für Prosodie, Aussprache, und Hypernasalität genutzt. Ein weiterer Ansatz dieser Arbeit ist die Verwendung eines Kindersprache Erkenners zur Modellierung eines naiven Zuhörers. Wenn die Aufnahme Bedingungen konstant gehalten werden sollte der einzige variierende Faktor der Mensch sein. Daher kann die Erkennungsrate die Verständlichkeit des Sprechers repräsentieren.

Die Sammlung von Patientendaten erfolgte in Erlangen von 2002 bis 2008. 312 Kinder mit LKG wurden aufgezeichnet. Kontrollgruppen wurden in vier großen Städten in Deutschland gesammelt um mehrere regionale Dialekte abzudecken. 726 Kontrolldatensätze wurden aufgezeichnet.

Die experimentellen Ergebnisse zeigten, dass das automatische System hohe und signifikante Übereinstimmungen zu den menschlichen Bewertern für globale Parameter wie Verständlichkeit sowie einzelne Artikulationsstörungen hat. Das System ist in der gleichen Größenordnung wie die menschlichen Bewerter. Die Verständlichkeitsbewertung erwies sich als unabhängig von dem regionalen Dialekt. Die Visualisierung der Sprachdaten zeigte auch eine hohe Zustimmung zu subjektiv bewerteten Kriterien. Artefakte, die durch die Verwendung mehrerer Mikrofone entstehen, wurden entfernt.

Contents

Chapter 1

Introduction

Automatic speech recognition has become a popular and wide-spread technology. Current state-of-the-art systems work fast and reliably. These techniques proved to be applicable in many working scenarios:

- **Dialogue Systems:** When someone calls e.g. his insurance company, one often gets connected to an automatic dialogue system. The system asks the user for the reason of his call and connects him to the desired department. So the call center agents only face questions of their expertise which saves the company a lot of time and work. Since some systems are also able to connect a phone call directly to a certain person, some of the agents even prefer to use the system as directory assistance instead of their telephone book [Haas 07].

- **Mobile Phones:** Cellular phones often use speech recognition to improve the comfort in dialing. State-of-the-art phones don't even have to be trained or adapted to its user because they are already shipped with a speaker-independent speech recognizer.

- **Dictation Systems:** Especially in professions where it is necessary to create a lot of correspondence dictation systems alleviate the work a lot. Lawyers and medical doctors often use personalized speech recognizers to write their reports.

- **Voice Command Systems:** In the automotive sector, so-called *hands-free* voice command systems are used to control different appliances, e.g., navigation systems. So the driver is not required to remove his hands from the steering wheel.

In medical applications the use of automatic evaluation and expert systems increased rapidly over the last decades. One of the first expert systems was MYCIN [Bucha 84] which was developed at Stanford University. With a rule set of about 500 rules it could diagnose infectious blood diseases. By asking the user a long set of textual yes/no questions, it could analyze the type of the bacteria and recommend antibiotics.

Nowadays medical systems are much more advanced. A currently running project — *Health-e-Child* [1] — aims at developing an integrated health care platform for European pediatrics. Its goal is to provide the physician access to biomedical knowledge

[1] founded by the European Union under grant IST-2004-027749

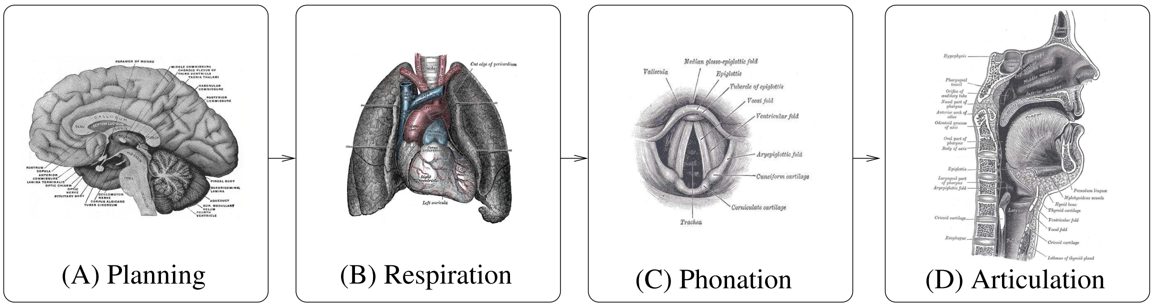


Figure 1.1: Simplified scheme of human speech production (engravings from [Gray 18])

repositories in order to enable him to compare the case he is currently investigating with similar ones.

Surprisingly, speech recognition techniques are rarely used in medical contexts. In most of the cases, only dictation systems are used. In this work we show that speech recognition yields much more than this: If speech recognition techniques are applied in controlled conditions, they can be used to qualify and quantify different properties of a person's speech.

1.1 Motivation

Communication is important for our daily life. About 87.5 % of the inhabitants of urban areas require communication for their daily work. Communication disorders cause a major effect on the economy. The cost of care as well as the degradation of the employment opportunities for people with communication disorders cause a loss of $154 billion to $186 billion per year to the economy of the United States of America alone. This equals to 2.5 % to 3.0 % of the Gross National Product of the US. These facts indicate that communication disorders are a major challenge in the 21st century [Ruben 00]. The use of automatic speech processing techniques will contribute to reduce the cost of the care of communication disorders as well as provide better rehabilitation of such disorders and hence increase the employment opportunities for people with such disorders.

Communication between two persons — the sender and the receiver — is a complex process. It is only possible if both communicating sides share the same language, i.e., know syntax and morphology of the language. Human languages are assembled from phonemes which are the smallest structural units that distinguishes the meaning. An example for a phoneme is /i:/ as it appears in the word “b**ea**t” and “m**ee**t”[2]. A sequence of phonemes forms a word. The word “beat” is constructed of the phoneme sequence /b i: t/. In turn, sequences of words are used to build sentences.

Figure 1.1 shows a basic scheme of speech production: First the speech is planned in the brain (A). Here, the information which is to be transmitted to the receiver

[2] Note that the phoneme-grapheme relation is not one-to-one and ambiguous.

is translated into sentences and words. Then a plan has to be created to articulate these words, i.e., to code the words as a sequence of phonemes.

In order to pronounce the words, air is emitted from the lung (B) through the larynx. The vocal cords (C) can either be opened widely to produce unvoiced speech, or closed in order to create voiced speech. In closed condition the vocal folds oscillate when air streams through them. This produces the primary voice signal — fundamental frequency and the harmonic structures.

In the vocal tract (D), finally the phonemes are formed by the human articulators, i.e., lips, teeth, jaw, tongue, and palate. In source-filter theory [Fant 60a] the vocal tract is modeled as a tube which consists of segments with different diameters. Depending on the position of the articulators, the diameters of the tube segments vary and hence also the filter, i.e., the impulse response of the linear system of tube segments. The actual articulation is then modeled as a convolution of the source signal, i.e., the fundamental frequency and the filter response of the vocal tract. Filtering of the source signal forms the final speech signal as it is emitted from the mouth. Different positions and motions of the articulators result in distinct speech samples which can be interpreted as phonemes by other listeners. Hence, each phoneme corresponds to characteristic movements of the articulators.

Speech production is age- and gender-dependent in general. On the one hand this is related to anatomical differences. In adults, the vocal cords are between 12.5 mm and 17.5 mm in length for females and 17 mm to 25 mm for males. The resulting average fundamental frequencies are, therefore, about 125 Hz in males and 210 Hz in females. In children fundamental frequencies above 300 Hz occur. Also, the length of the vocal tract changes from about 10 cm in children to 17 cm in adults. On the other hand, children still acquire speech, i.e., it is a known fact that children's speech intelligibility is connected to the age of the children. The older the children, the higher is their speech intelligibility [Wilpo 96].

In communication disorder theory, three different types of disorders are distinguished: language, voice, and articulation disorders. Each of these disorders originate from a different place in speech production.

Language disorders are caused by the first element of the speech production chain (cf. Figure 1.1 (A)) — the brain. Such a disorder (e.g. aphasia) can be caused by a brain injury, a stroke, tumors, or other cerebral diseases. It affects auditory comprehension, oral-expressive language and reading. These patients may have a markedly reduced auditory comprehension, are unable to repeat words, and are not able to name items. The oral-expressive language might be sparse or unintelligible [Wertz 04].

Language disorders in children are very heterogeneous. They range from expressive difficulties to severe receptive disorders. Reading and writing disorders may accompany language disorders. In general, language disorders, learning disorders, attention deficit disorders, and disruptive behavior are well-known to coincide [Winds 04]. Reading disorders, for example, can be predicted from early difficulties in the expressive language or the comprehension [Catts 97].

Voice disorders occur if either the breathing or the phonation is affected (cf. Figure 1.1 (B/C)). Severe voice disorders can result from the removal of the larynx. After the larynx was removed, the trachea is detoured to an opening in the throat (tracheostoma). So the patient is able to breath after surgery. In order to enable the patient to speak again, a one-way shunt valve is placed between the trachea and the esophagus. So the patient can speak by breathing in, closing the tracheostoma, and breathing out through the esophagus. This results in an extremely rough voice which sounds like the voices of early speech synthesis [Schut 02, Hader 06a].

The cause of articulation disorders is the last element in the chain of speech production (cf. Figure 1.1 (D)). Such problems can arise from structural changes of the vocal tract. For the case of children with cleft lip and palate (CLP) — who are investigated in this work — the main causes are enhanced nasal air emission which leads to altered nasality, a shift in localization of articulation (e.g. using a /d/ built with the tip of the tongue instead of a /g/ built with the back of the tongue or vice versa), and a modified articulatory tension [Hardi 98].

1.2 State-of-the-art Diagnostics in Voice and Articulation Disorders

Each of these disorders is evaluated with a different assessment scheme. For the evaluation of language disorders, the subject usually has to produce speech, e.g. in standardized interviews. Voice disorder evaluation relies mainly on sustained vowels. In order to identify the properties of an articulation disorder, the subjects have to pronounce certain words which cover all phones of the respective language. Except for voice disorders, most of the evaluations are done subjectively. Thus, the results can hold large differences between different experts [Paal 05].

In order to attenuate the differences between multiple experts, the mean of their opinion is usually formed for scientific purposes. This inter-subjectively verified mean is then often called objective in the literature. However, it is still obtained subjectively and contains interpretations and influences. Nevertheless, this is the only method that can create a "gold standard" which is not just dependent on a single opinion if no objective methods are available.

In clinical practice, however, it is not possible to do all evaluations with a panel of experts. Therefore, just a single expert assesses the disorder. This method works quite well as long as the therapist sees the patient regularly and the therapist does not change. However, it is difficult to compare the subjective evaluations. Thus, if panels of experts are too time-consuming and single experts are not reliable, an objective evaluation method is required.

In speech therapy only few objective evaluation methods are known. Most of them are based on the examination of sustained vowels, i.e., jitter and shimmer computed on these vowels. For the dysphonia severity index (DSI) [Wuyts 00], for example, the subject has to pronounce an /a:/ as long as possible. Furthermore, he has to produce a frequency as high as possible and a vowel as quiet as possible. The severity is then computed from the maximum phonation time, the lowest phonation energy, the mean jitter, and the highest fundamental frequency which was reached.

Recently, a novel objective method for the assessment of the intelligibility was presented by our group [Hader 04] for the substitute voices of laryngectomees. It is the first objective method which analyses continuous speech. Previous automatic methods only evaluated sustained vowels. The idea is to use a word recognizer instead of a speech therapist as “listener”. The recognizer was trained with normal speech of adults from all dialect regions of Germany. As shown in [Stemm 05], the recognizer produces a stable recognition with a small — but not negligible — error rate. When the recognizer is confronted with a severe voice disorder, the recognition rate drops dramatically. If the test is done in controlled conditions and all other causes for bad recognition like noise or a bad microphone are eliminated, this effect can be used to obtain an estimate for the intelligibility of the test subject.

1.3 Contribution of this Work

In this work we extend the approach from [Hader 04] to assess the speech of children with CLP. This task is more difficult than the analysis of adults’ voice disorders. In comparison to adults’ speech, the speech of children shows a much wider spectral variability which makes speech recognition more difficult in general [Wilpo 96, Lee 97, Li 01].

During this work a client-server platform to analyze the children’s speech data was created. Due to the high demand from the medical cooperation partners, soon more studies were performed using our system. Therefore, it is now called “Platform for Evaluation and Analysis of all Kinds of Speech” (PEAKS) [Maier 07b].

The client runs in any web browser with Java support without any installation routines. It is platform-independent and was tested on Linux, Windows, and Mac operating systems. So the therapists can record patients’ data with a standard PC which is connected to the Internet without any modifications or upgrades. In order to get a similar audio quality with the different PCs, we use USB headsets which have their own sound card attached. The data are digitized by the headset which ensures a comparable quality on any PC.

In order to enable recording without Internet, a portable version of PEAKS was created as well. Therefore, a local server which stores the recordings until the computer is connected to the Internet again is installed on a portable computer.

The Server does all the evaluations and analyses. It is installed on a Linux system and runs the analysis tools of the Chair of Pattern recognition which have been developed there for the last 20 years. As soon as a client transmits data to the server, the data is stored and different analyses can be requested by the client. The server supports three main types of analyses:

- intelligibility measurement,
- articulation assessment, and
- visualization.

Using these analyses the speech data can be evaluated repeatedly which enables a tracking of the patient’s development during the speech therapy.

1.3.1 Intelligibility Measurement

The first main topic in this work is the automatic measurement of the intelligibility of children's speech. Usually, the children are shown pictograms whose names have to be uttered in order to test their speech. The evaluation of children's speech is much more difficult than the measurement of the intelligibility of adults' speech due to the fact that the automatic recognition is more complex and that the reference of the test are pictograms. These pictograms can be misunderstood easily: e.g. the pictogram of the "hunter" is often named "a man with a dog". Thus, the therapist has to correct the child. Another reason for the therapist to interfere is that children with articulation disorders often do not want to talk. Then the therapist has to encourage the child to say the target words. To compensate these effects we apply several techniques:

- **Speaker recognition:** In order to enable the therapist to interfere during the speech test, the child's audio data and the therapist's data have to be separated automatically. This is done with speaker recognition techniques: Directly after the registration of a new therapist, the system asks the therapist to perform a recording of his own speech. With these data a speaker model is generated. Using the model and a background model for various children, the data of each speech test can be separated.

- **Speaker adaptation:** Children's speech shows a high variablity per se. This effect has to be compensated. Adaptation techniques have shown to increase recognition performance and robustness in various difficult acoustic conditions [Maier 05a, Gales 96]. In this work we show that it improves the quality of the intelligibility measurement as well.

- **Word spotting:** The setup of the speech test is designed to analyze only the words which appear in the test. Therefore, word spotting has to be applied to recognize just the words which are relevant for the speech test. Analyses would be distorted if additional words would be analyzed by the system.

- **Speech recognition:** In order to model a "naive" listener, we apply a speech recognizer which is trained with speech data of children without CLP. Since the words which should be uttered by the patient are known a priori and all the test conditions are kept equal, a recognition rate can be calculated. This rate represents the intelligibility of the speaker because it is the only factor that is changed during the recordings. So we can use standard evaluation methods for speech recognizers to assess the intelligibility of a patient.

- **Prosodic feature extraction:** The human rating of intelligibility is dependent on the speaking style. Children who speak faster and more vivid are perceived as better intelligible although the recognition performance is equal to other children with slightly lower intelligibility. To compensate this effect the prosodic information is included to the experts' rating estimation procedure.

By combination of these techniques we estimate a global mean of experts' intelligibility score. Furthermore, we predict the intelligibility scores of the individual experts with the analysis methods shown above.

1.3.2 Articulation Assessment

A more refined analysis of certain properties of the speech data is the articulation assessment. Its goal is to find defected phones or typical misarticulations which are not properly articulated by the patient. This can be performed on different levels of detail. Note that the features of the more detailed levels can always be used on a less detailed level by feature transformations like averaging over the respective time domain.

- **Test level:** The whole test is used for the analysis. This yields scores for different properties of the patient's speech. This procedure is similar to the evaluation of the intelligibility.
- **Word level:** Single words are analyzed in order to find particular speech problems. Therefore, a list of pathologic and non-pathologic alternatives of the test's words has to be generated. These words are then added to the recognizer's vocabulary. The generation of this list can be done by a rule set [Hessl 05] or a data-driven method which analyses the transliteration of the data. Furthermore, this can be combined with pronunciation features [Hacke 05a, Hacke 05b]. These features are designed to model the pronunciation independent of the language. Basically these features represent the difference between correctly and incorrectly pronounced words. Although they were designed for the assessment of non-native speech they can be used to assess the speech of children with cleft lip and palate as well because the features were designed to model the deviation from "normal" speech.
- **Phone level:** By application of a phone recognizer, the evaluation is done for each phone individually. Similar pronunciation features like on word level are computed for each phone. Although the classification rates are low these features improve the classification process on word level.
- **Frame level:** A frame is a short time segment of speech data. For the case of this thesis, its length is 16 ms. This is about one fifth of the duration of an average phone. A decision for the class of each frame is done every 10 ms. Although the time is short the combination of many decisions achieves good results on lower levels of detail.

The articulation assessment is used to identify probable articulation problems of a patient. For this automatic testing procedure, there is no need for a speech therapist to be present. The test could also be performed by an assistant who was instructed how to use the system. In this manner screening of children at a certain age, e.g. before they enroll in primary school, could be performed in order to detect children with articulation problems. If such a problem is detected by the automatic system, it is advisable to consult a speech therapist whether further therapy is necessary. The intention of PEAKS is to create a diagnostic tool rather than to replace the therapist[3].

[3] Of course one can estimate the body temperature of a patient just with his own hand. However, in modern medicine the use of a thermometer has prevailed.

1.3.3 Visualization

Furthermore, we show a method to visualize speech disorders. The idea is to create a 2-D or 3-D map which shows well-documented patients and is used to compare new patients to the well-documented ones. This gives the medical personnel a better understanding of the articulation disorder and how it is related to others. The goal of our research is to find a method which maps the patient into a space with known regions. These regions should represent the properties of the patients' speech. So an "articulation disorder space" is generated. One of the most important features of this space is that patients with similar articulation disorders are mapped close together. Therefore, we introduced a mapping method [Hader 06b] to visualize differences between different speakers for voice disorders. This mapping method projects the parameters of a speaker-adapted ASR to a 2-D or 3-D space which can be visualized easily. Due to the large number of parameters, we chose a nonlinear dimension reduction method which preserves the topology of the high-dimensional space. The method was applied to the children's speech data of this work. In the resulting map, meaningful regions were found which are used to describe the voice disorder in detail.

1.3.4 Tracking of the Therapy's Progress

All of the different methods to analyze articulation disorders are used to track the progress of the therapy. We show that the intelligibility of children increases with their age. Age-dependency is a very important attribute which has to be taken into account in the system. If age-dependent values are supplied in normal children, the system can be used by speech therapists to evaluate different methods of treatment. Thus, the best method to handle each articulation disorder can be found. This, however, is not topic of this thesis.

1.4 Outline of this Work

This work is organized as follows: The second chapter deals with cleft lip and palate. The origin of the structural changes is described as far as it is known today. The consequences of these changes are mentioned as well. The primary closure of the cleft lip or palate is often performed in the age from 6 to 15 months. However, further treatment is often necessary in order to improve the quality of speech in the children. At the end of the chapter, the effects of the clefting on speech production and their acoustic properties are outlined.

In the third chapter, the state-of-the-art of the evaluation of disordered speech is discussed. Most of the evaluations are done subjectively. Thereby, different strategies can be observed: Some raters give very detailed evaluations where they focus on aspects while others rather use holistic impressions as their parameters. Often averaging over multiple labelers helps to reduce the subjectivity of each rater in scientific contexts. However, this is not feasible for clinical practice due to the high costs. Another way to omit the problem of subjectivity is to use many naive raters instead of a few experts. Thus, the problem arises that some properties are too specific to be rated by naive listeners.

A solution is the application of objective methods. However, only few objective measurements exist in speech and voice therapy. We present the objective methods for hoarseness, nasality, and intelligibility. Most of these methods require special expensive hardware, are too specific, or cause other problems which makes them infeasible. Chapter 3 is concluded by the description of a method to determine a proper ground truth which is used throughout this thesis.

The beginning of Chapter 4 introduces the PEAKS system. First the requirements of such a system are specified. The system setup describes data and process management, data acquisition, preprocessing, feature extraction, assessment, report generation, and where these processes take place. Furthermore, the system's collaboration and security concepts are pointed out.

The following sections contain the mathematical and algorithmic concepts. In order to compare the speech evaluation by different raters, their agreement has to be measured. This is done with regression, correlation analysis, and Kappa and Alpha coefficients of agreement. For assessment by the automatic system, support vector machines, support vector regression, and Gaussian mixture models (GMMs) are used in this work. The dimensionality of a vector space is reduced by principal component analysis, linear discriminant analysis, and the Sammon mapping. Speech processing techniques from feature extraction over speaker recognition, acoustic modeling, language modeling, decoding, prosodic analysis, pronunciation analysis are explained as well. The section is concluded by different techniques for the normalization of age effects.

All modules of PEAKS are presented in Chapter 4.3. Preprocessing in PEAKS is mostly speaker segmentation. The features which are extracted by the PEAKS system are word accuracy, word correctness, prosodic features, and pronunciation features. Classification is done for the intelligibility assessment and the articulation assessment. In the end a report with a visualization of the result is generated.

The fifth chapter presents the databases which were used in this work. First the PLAKSS test which was used for all recordings in this work is described in detail. Then the recordings of normal children which form the control group of this work are presented. A description of the speech data of the patient groups which were collected during this work follows. Chapter 5 ends with the specification of the training speakers of the recognizer.

Chapter 6 contains the results of the experiments which were done to substantiate the theories of this work. We present that the speaker segmentation works properly. This is followed by an elaborate documentation of the experiments on intelligibility assessment. Next, the results on articulation assessment are presented. The chapter is concluded by visualizations of articulation disorders.

At the end of this thesis, Chapter 7 gives an outlook on future work, and Chapter 8 summarizes the most important aspects of this work.

Chapter 2

Cleft Lip and Palate

In this chapter the malformations of the palate and the lip, their genesis, and the effects on the patient are described. Furthermore, the surgical treatment and the subsequent therapy according to the interdisciplinary *Concept of Erlangen* [Wohll 04, Bautz 08] is introduced. The last section of this chapter states the effects on speech production and their acoustic properties.

2.1 Epidemiology, Etiology, and Functional Consequences

Orofacial clefting shows a broad spectrum of different clefts [Epple 05]. An anatomical classification was done by Tessier [Tessi 76]. He assigned the numbers from 0 to 30 to the positions of the different cleavages. Figure 2.1 displays only the most important types (numbers 0 to 14 and 30). These types are:

- orofacial clefts (numbers 0 to 7, 30)
    - median cleft (number 0)
    - unilateral and bilateral clefts (i.e., cleft numbers 1, 2, and 3)
    - oblique facial clefts (numbers 4 and 5)
    - lateral facial cleft (numbers 6 and 7)
    - median mandibular cleft (number 30)
- craniofacial clefts (numbers 8 to 14)
- rare clefts (numbers 15 to 29; not included in Figure 2.1)

In this work we focus on unilateral cleft lip (UCL), the bilateral cleft lip (BCL), the cleft palate (CP), or their combinations — unilateral cleft lip and palate (UCLP) and bilateral cleft lip and palate (BCLP) — denoted with numbers 0 to 3 in the scheme of Tessier. The terms *cleft lip* (CL) or *cleft lip and palate* (CLP) are used to characterize the union of the respective unilateral and bilateral groups.

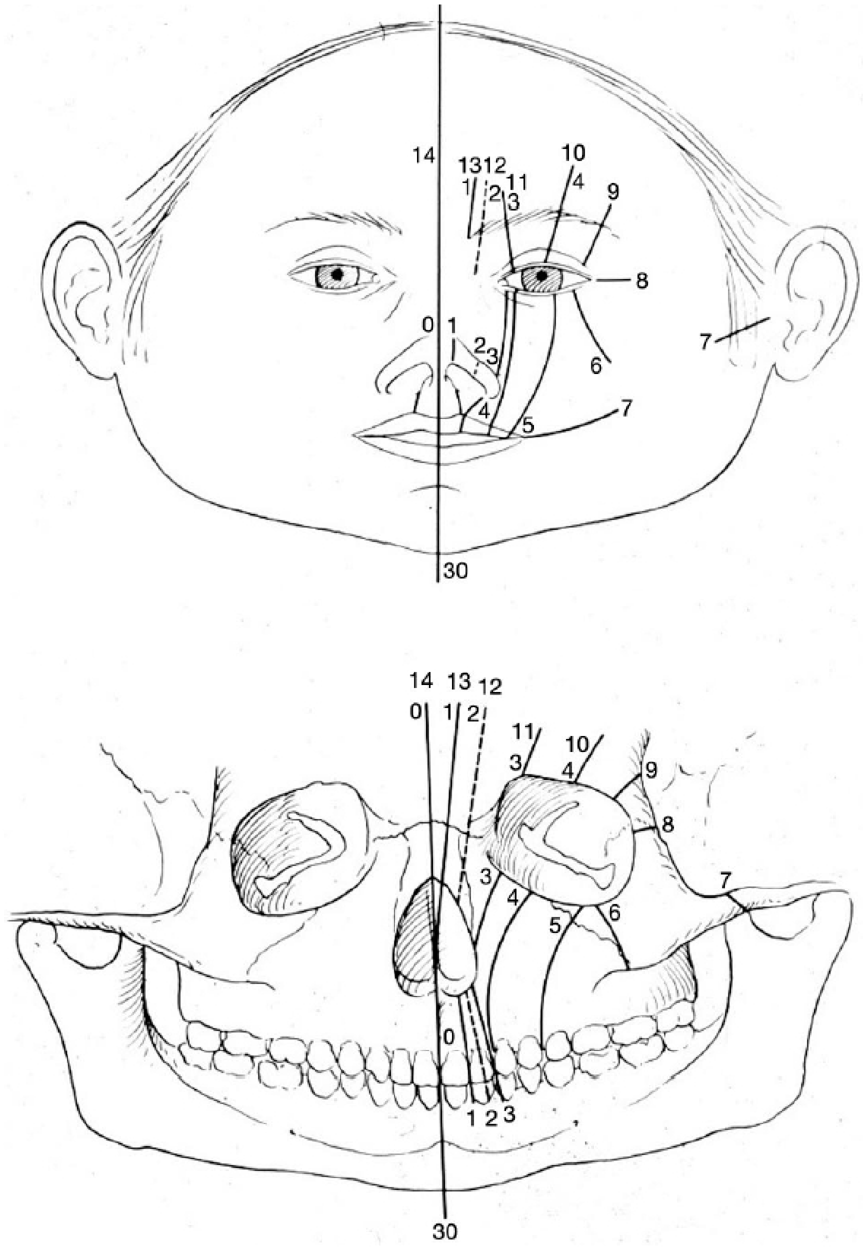


Figure 2.1: Classification of different cleft types according to Tessier [Tessi 76] (Figure taken from [Epple 05]).

2.1.1 Epidemiology

CLP is the most common malformation of the head. It constitutes almost two thirds of the major facial defects and almost 80 % of all orofacial clefts [Epple 05]. Its prevalence differs in different races from 1 in 400 to 500 newborns in Asians to 1 in 1500 to 2000 in African Americans. The prevalence in Caucasians is 1 in 750 to 900 births [Tolar 98, Kawam 90]. Clefts on the left side are—for reasons that are yet unclear—more often than on the right side [Epple 05].

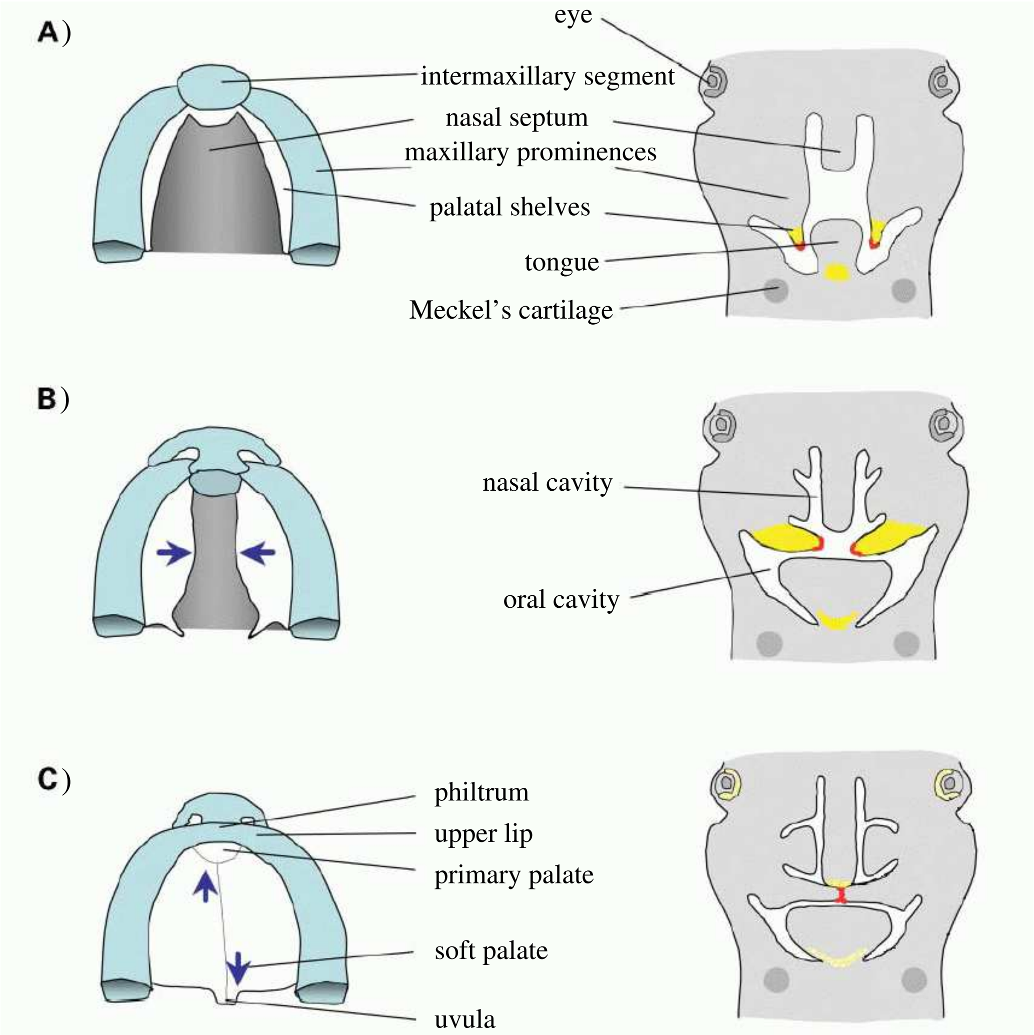


Figure 2.2: Stages of the embryonic development of lip and palate: A) illustrates the paired horizontal and coronal sections of the head at week 7. B) shows the development until week 8. The palatal shelves elevate to a horizontal position above the tongue. C) shows the development until week 9–10: the palatal shelves fuse together with the nasal septum to form the primary and soft palate (adapted from [Stani 04]).

2.1.2 Embryology and Etiology

The lip and the palate develop in the human embryo from week 7 to week 10. Figure 2.2 displays the process in three stages. In week 7 (Figure 2.2 A) the medial nasal prominences join to form the inter-maxillary segment. Until week 8 (Figure 2.2 B) the palatal shelves lift and begin to move to the center. The arrows mark the initial position of contact of both shelves. In week 9–10 the fusion of the palatal shelves is complete. The anterior part forms the primary palate while the posterior part develops to the soft palate and the uvula [Stani 04].

Basically all types of CLP are caused by an insufficient contact or fusion of either the palatal shelves (cleft palate) or the maxillary segment (cleft lip). It is suspected that besides the merging of the palatal shelves and the maxillary segment the size of the facial processes affects clefting as well. Races with a rather small median nasal process like Asians have therefore a higher tendency to clefting than African Americans [Epple 05].

The causes of CLP are manifold and include genetic and environmental factors. In 15 % of the CLP cases, the clefting is part of a syndrome. In fact, there are 171 syndromes which involve CLP. However, transmission of the cleft phenotype in a Mendelian manner occurs seldom. By now the molecular basis of human clefting was not found. Nevertheless, some genes which might contribute to clefting were identified [Epple 05, Stani 04]. The list of environmental factors which increase the chance of clefting is growing. By now the following factors were identified [Carin 03] to increase the risk of facial clefting:

- Alcohol
- Cigarette smoke
- Inappropriate nutrition
- Steroids
- Anticonvulsants[1]
- High altitude

2.1.3 Functional Consequences

CLP can result in morphological and functional disorders [Wanti 02] whereat one has to differentiate primary from secondary disorders [Milla 01, Rosan 02]. Primary disorders include problems of nutrition, swallowing, breathing and mimic disorders due to the clefting. After surgical closure of the cleft, most of the primary disorders are corrected. Speech and voice disorders [Schon 94, Lierd 03] as well as conductive hearing loss caused by insufficient aeration of the middle ear[2] [Palio 05, Schon 99] are secondary disorders. Speech disorders can still be present after reconstructive surgical treatment.

The characteristics of articulation disorders are mainly a combination of different articulatory features, e.g. enhanced nasal air emissions that lead to altered nasality, a shift in localization of articulation (e.g. using a /d/ built with the tip of the tongue instead of a /g/ built with back of the tongue or vice versa), and a modified articulatory tension (e.g. weakening of the plosives /t/, /k/, /p/) [Hardi 98]. They affect not only intelligibility but therewith the social competence and emotional development of a child.

In the literature many examples are found. However, the number and kind of the reported features is vast and inconsistent in some cases. Therefore, we will first give an overview on the literature and summarize the structure of the speech of children with CLP in the end.

[1] anti-epileptic drugs to suppress the excessive firing of neurons

[2] The clefting and its closure might influence the function of the Eustachian tube. Hence, pressurization might not be possible and the hearing ability is reduced.

Study	# CL	# CP	# UCLP	# BCLP	# CONTR	$\sum$ /
[Karli 93b]	-	-	84	19	40	143
[Peter 95]	-	11	53	46	-	110
[Ysunz 97]	-	15	-	-	-	15
[Laiti 98]	82	82	85	31	-	280
[Bress 99b]	-	1	92	35	-	128
[Schon 99]	16	96	81	77	-	270
[Laiti 00]	49	34	33	17	-	133
[Pampl 00]	-	58	-	-	-	58
[Lierd 01]	-	-	2	-	-	2
[Nakaj 01]	-	33	74	28	168	303
[Pulkk 01]	-	35	30	-	-	65
[Sell 01]	-	-	647	-	-	647
[Timmo 01]	-	27	17	-	-	44
[Young 01]	-	-	38	-	-	38
[Bress 02]	-	-	-	124	-	124
[Gibbo 02]	-	27	-	-	-	27
[Lierd 02]	-	-	19	18	54	91
[Lohma 02]	-	22	-	-	-	22
[Pulkk 02]	38	33	44	19	-	134
[Lierd 03]	-	-	8	6	n/a	14
[Morri 03]	-	20	-	-	-	20
[Schus 03]	-	21	30		-	51
[Lierd 04]	103				-	103
[Ysunz 04]	-	-	70	-	-	70
[Brunn 05]	-	11	-	-	-	11
[Palio 05]	-	9	19	14	-	42
[Pampl 05]	-	90	-	-	-	90
[Laiti 06]	-	-	17	-	17	34
[Willa 06]	-	-	38	-	36	74

Table 2.1: Studies on differences in the characteristics of speech of children and adolescents with cleft lip (CL), cleft palate (CP), unilateral cleft lip and palate (UCLP), and bilateral cleft lip and palate (BCLP); The table reports the number of investigated subjects. Some studies compare the speech to a control group (CONTR).

In [Karli 93b] it is shown that the speech of children with unilateral cleft lip and palate (UCLP) and bilateral cleft lip and palate (BCLP) is poorer than the speech of the children in the control group (CONTR) with respect to hypernasality, intelligibility, nasal escape, hyponasality, and other deviant articulation caused by CLP. Speech of children with BCLP is significantly poorer than speech of children with UCLP in intelligibility and other deviant articulation caused by CLP. [Lierd 02] reports that UCLP and BCLP speech is worse than CONTR in intelligibility, hypernasality, nasal emission, nasalance measured with a nasometer [Kay 94], and the mirror fogging test. In contradiction to [Karli 93b], no significant differences between UCLP and BCLP are reported in this study which might be related to the number of tested subjects (cf. Table 2.1). In [Lierd 03] the results of [Lierd 02] are confirmed with even fewer subjects. Furthermore, the voice quality is measured with the Dysphonia Severity Index (DSI) [Wuyts 00]. No significant differences between UCLP and BCLP are found. The voice quality measured with DSI is normal or just slightly reduced in the CLP children.

Another feature of speech of children with cleft lip and palate is compensatory articulation and misarticulation. They are caused by anatomic deficits [Pulkk 02]. In order to compensate these deficits, the children learn to form similar but yet different phones which can result in a chronic speech disorder. It was shown that this malformed articulation results in a lower rate of speech [Bress 99b]. Double articulations, i.e., consonants which are articulated with two places of articulation simultaneously, appear rarely in CLP children [Gibbo 02].

The occurrence of misarticulations of Finnish dental consonants (/r/, /s/, and /l/) was reported to be maximal in children with bilateral cleft lip and palate while children with just cleft lip had the lowest number of misarticulations [Laiti 98]. Furthermore, it was shown that the elimination of misarticulations of /s/ and /l/ occurs more often in the age between 6 and 8 years than the elimination of /r/ misarticulations in Finnish Children [Laiti 00].

[Sell 01] presents a perceptive evaluation of the nasality, the intelligibility and articulation errors. Minor and serious consonant errors were annotated (cf. Table 2.2). Other consonant errors are rather uncommon according to [Sell 01]. Furthermore, the intelligibility, the nasality, and the number of consonant errors of 5-year-old children are worse than those of 12-year-olds.

[Pampl 00] compares CP children with and without compensatory articulation disorders. The children with compensatory articulation disorders have a significant delay in language development. [Morri 03] states a comparison between two groups of CP children at two and three years: The first group had significantly delayed language development while the second had no delay. The delayed language development persisted at the age of three years in the first group.

2.2 Treatment

All children of this study were treated according to the *Concept of Erlangen* [Wohll 04]. The concept states a full interdisciplinary treatment of the children including maxillofacial surgery, orthodontics, phoniatrics and pedaudiology, oto-rhino-laryngology, pediatrics, gynecology, geneticists, speech therapy, and parental counseling. As Ta-

articulation errors	
minor errors	explanation
lateralization	tongue is in a lateral position between the teeth during articulation
palatalization	tongue is touching the palate during articulation
interdentalization	tongue is between the front teeth during articulation
serious errors	example
pharyngealization	tongue is shifted backwards towards the pharynx during articulation
glottal articulation	the closure of the plosives is done in a glottal manner instead of a labial. This is also called laryngeal replacement in the following.
backing to uvular	the tongue is shifted backwards towards the uvula
backing to velar	the tongue is shifted backwards towards the soft palate
active nasal fricatives	air is emitted though the nose during the articulation of fricatives
absent pressure consonants	plosives are not formed or weakened during the articulation
nasal realizations	the nasal air flow is persistent throughout the whole word
weak nasalized consonants	the consonants are nasalized, i.e., air is emitted during the articulation of the consonants

Table 2.2: Minor and serious articulation errors after [Sell 01]

<table>
<tr><th colspan="2">Concept of Erlangen</th></tr>
<tr><td>first few hours</td><td>A palatal obturator is used to occlude the clefting. This helps in the nutrition and the breathing of the child.</td></tr>
<tr><td>first few days</td><td>The parents are counseled and informed about orofacial cleftings.</td></tr>
<tr><td>6th month</td><td>The clefts of the lip and the mucosa of the hard palate are closed. A tympanostomy tube is inserted into the eardrum in order to aerate the middle ear if required. The hearing ability of the child is tested.</td></tr>
<tr><td>10–15th month</td><td>The soft palate is closed and, if necessary, a tympanostomy tube is inserted into the eardrum.</td></tr>
<tr><td colspan="2">End of the primary surgical treatment
Follow-up examinations every year</td></tr>
<tr><td>7th year</td><td>Begin of the orthognathic surgery: Closure of remaining clefts between the oral and the nasal cavity</td></tr>
<tr><td>10th year</td><td>Correction of the jaw (osteotomy)</td></tr>
<tr><td>12th year</td><td>Alignment of the upper and the lower jaw if necessary</td></tr>
<tr><td>18th year</td><td>Secondary osteotomy and further speech-improving interventions if necessary</td></tr>
</table>

Table 2.3: Chronological overview of the *Concept of Erlangen* [Wohll 04]

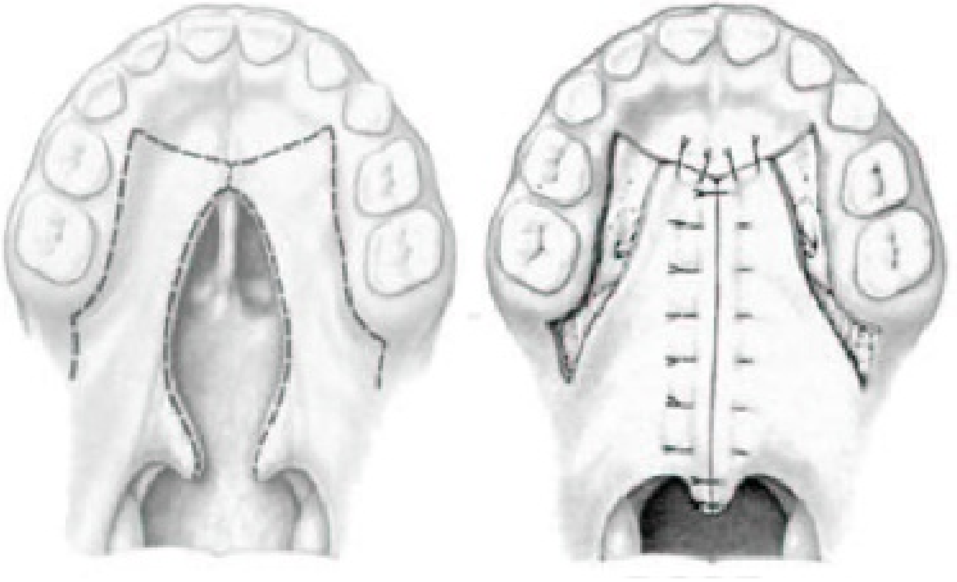


Figure 2.3: Closure of the palatal cleft: The hard and soft palate are sewed together [Hausa 00].

ble 2.3 shows, this concept of treatment starts right from the birth. Just a few hours after the mother gives birth to a child with CLP, a palatal obturator is inserted into the mouth of the infant. This prothesis occludes the clefting and supports nutrition and breathing. In the following days, the parents are further counseled about the disease of their child [Young 01]. After adequate information of the parents, most feel less stressed compared to parents whose children have other handicaps [Schus 03]. With consistent team care including follow-up examinations, treatments, and assistance to the family, the therapy is alleviated a lot [Peter 95].

Usually the cleft palate children have their primary surgeries within the first two years of their life. Since multiple attempts on the reconstruction of the palate result in negative effects on the speech outcome [Bress 02], the number of interventions is as small as possible in the Concept of Erlangen. Because early closure might have a positive effect on the pre-linguistic development [Willa 06], the palatoplasty is performed early. The palatal cleft is closed as displayed in Figure 2.3. The closure of the soft palate follows at the age of 10 to 15 months. Often a tympanostomy tube is inserted into the eardrum in order to keep the aeration of the middle ear to prevent from hearing loss. This concludes the primary surgical treatment.

Speech therapy according to the *Concept of Erlangen* starts as early as needed, sometimes right with the birth of the child, e.g. with breastfeeding support. The therapy is adjusted to the individual needs of the child and the disorder.

In recent research the application of biofeedback training [Ysunz 97, Brunn 05] was also beneficial. Intensive speech therapy in a summer camp also yields positive effects in the treatment of articulation disorders [Pampl 05]. The use of the Internet to support an inexperienced speech pathologist by an experienced pathologist in his diagnosis is also tried in current research [Karne 05].

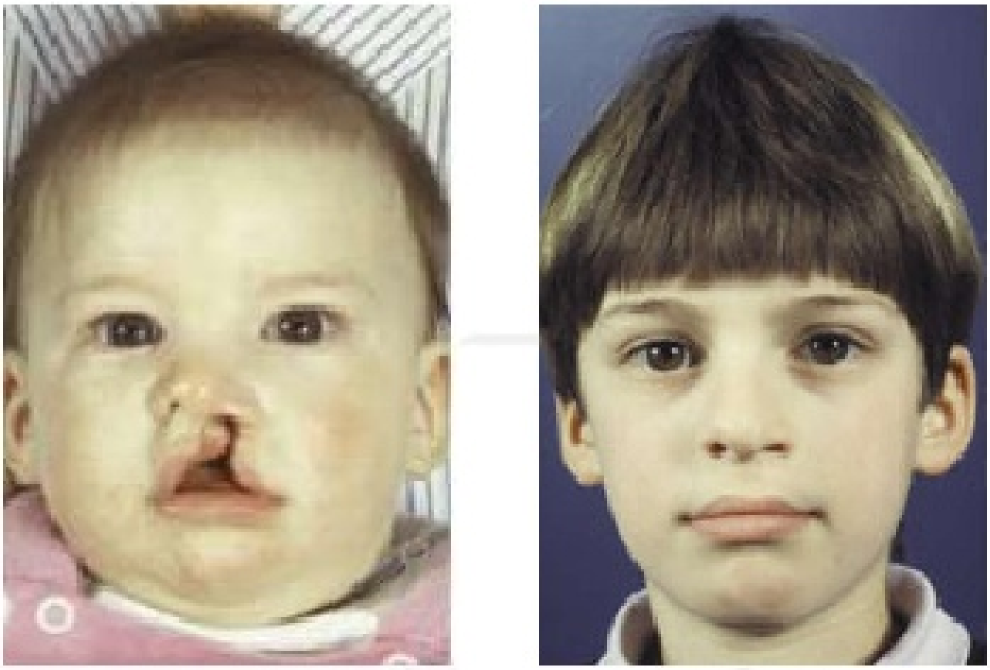


Figure 2.4: After adequate closure, the cleft (left side; age: 6 months) cannot be recognized anymore (right side; age: 9 years) [Hausa 00].

Sometimes, a second surgery is necessary to remove persistent hypernasality. This second surgery is independent of the method and timing of the primary palatoplasty [Pulkk 01]. A few years after the surgery, the cleft is barely recognizable (cf. Figure 2.4). Additional oral and maxillofacial surgery in the following years provides further corrections if necessary. Follow-up examinations at least once in a year are continued until the age of 18.

The interventions have many effects on the speech of these children. 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 [Lierd 04].

2.3 Summary

This chapter described the epidomology, the etiology, the treatment, the functional consequences, and the characteristics of speech of children with CLP. First, the development of the clefting during the different embryonic phases was portrayed. Then, the treatment—starting from the birth until the adolescence of the child—was explained in detail. The consequences of the cleft on nutrition, breathing, and especially on the speaking of the child were also reported. The effects on the speech of the children with CLP are various. As the most important aspects the following were recorded:

- With the perceptive assessment of the intelligibility and the measurement of the nasal airflow, significant differences between normal children and CLP children were measured [Karli 93b, Lierd 02, Lierd 03]. Significant differences in the intelligibility between unilateral CLP and bilateral CLP were found [Karli 93b] but could not be verified by all articles [Lierd 02, Lierd 03].

- The speech of the CLP children contains mild hyper- and hyponasality, dysphonia, and typical cleft type characteristics (interdentalization, lateralization, backing, glottal articulation, and absent pressure consonants) [Sell 01].

- No significant differences between isolated cleft palate and cleft lip and palate exist in the frequency of occurrence of certain articulation disorders. Furthermore, the speech outcome is similar in CP and CLP children [Timmo 01]. A relation between the size of the cleft and the speech deficits was found [Lohma 02].

- Delays in speech development [Pampl 00, Morri 03]—especially in children with bilateral cleft lip and palate [Nakaj 01] or syndromic forms of the disorder [Lierd 01]—and lower assertiveness in conversation [Laiti 06] were reported.

Chapter 3

Evaluation of Disordered Speech

This chapter is about the judging of speech, the quality and the characteristics of the "receiver". The task of evaluation and assessment of speech data from a certain speaker—called test subject in the following—is to assign a label which corresponds to a certain property of the speaker's speech. Basically, this process is the same for any criterion but different scales and evaluation schemes can be chosen. In the literature many methods to evaluate disordered speech are found. In general, these can be divided into two groups:

- perceptive evaluations which are performed by a human subject—also called rater or labeler in this context—and
- objective evaluations obtained by a device or algorithm which is independent of a human rater.

"Objective measurement" of speech in a sense that it is also independent of the test subject is not possible for the case of speech evaluation because the test subject has to utter a sequence of words or vowels in all cases. Therefore, objective measurement of speech disorders in this context could also be called "instrument-based".

For the perceptive evaluation, many different methods and scales can be applied. The two main methods are quantification and qualification. For the quantification of different properties of speech, direct magnitude estimation and equal-appearing interval scales are widely used. The qualification of certain characteristics of speech is often done by classification of small parts of speech like phones or words. Therefore, often classes like "present" and "not-present" are chosen. However, experienced raters can even distinguish further grades in such small parts of speech as discussed in Chapter 3.1.

Objective means exist only for quantitative measurements of nasal emissions [Kuttn 03, Lierd 02, Hogen 04] and for the detection of secondary voice or speech disorders [Bress 98, Zevce 02]. But other specific or non-specific articulation disorders in CLP as well as a global assessment of speech quality cannot be sufficiently objectively quantified. For the global assessment of speech intelligibility, just a single automatic approach is found in the literature [Schus 06a]. In Chapter 3.2 these objective approaches are presented.

Since an automatic system always needs to be trained with a "gold standard", the problem of the acquisition of such a reference is addressed at the end of this chapter.

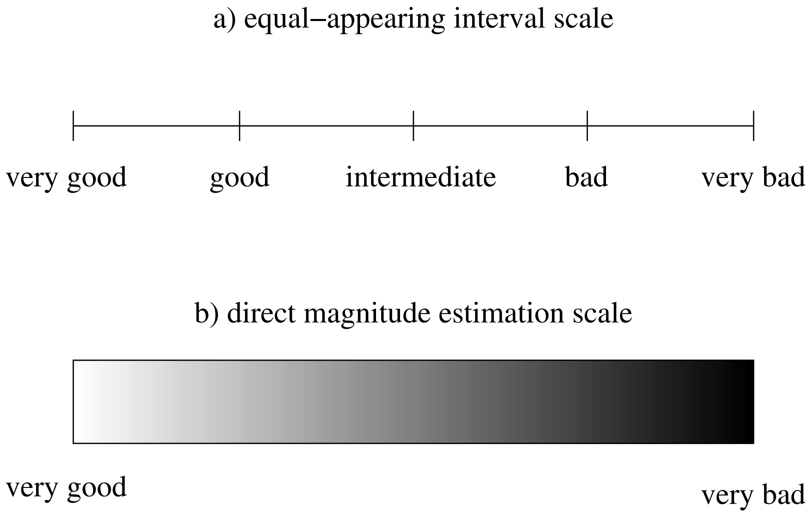

a) equal–appearing interval scale
very good
good
intermediate
bad
very bad
b) direct magnitude estimation scale
very good
very bad

Figure 3.1: a) shows a five-point equal-appearing interval scale with labels according to Likert [Liker 32] while b) shows a direct magnitude estimation scale as favored by [Schia 92].

There, the constraints which have to be met in order to create proper labels for an automatic system are discussed.

3.1 Perceptive Evaluation

The simplest method to evaluate speech disorders is auditive perception, often performed by speech therapists. For the qualitative assessment, the rater usually assigns certain nominal labels, i.e., classes to turns, words, or phonemes. Quantitative assessment, however, is more difficult: Usually, the listeners rate the test subject on an easy to understand scale. For this purpose two different kinds of scales are widely used [Schia 92]: equal-appearing interval scales and direct magnitude estimation scales (cf. Figure 3.1).

- **Equal-appearing interval scales:** They have the advantage, that they are very easy to understand and that raters get used to them very quickly. If such a scale is combined with nominal labels e.g. ranging from “very good” to “very bad”, the scale is called Likert-scale [Liker 32]. Table 3.1 a) shows an example for a five-point Likert-scale. Likert-scales are wide-spread and used in many assessment scenarios. In Germany, for example, all school marks follow such a scale. An absolute value is obtained for each test subject during the assessment procedure. Commonly used Likert-scales are the internationally used GRBAS [Hiran 81] and the German RBH [Wendl 05] scales for voice evaluation.

- **Direct magnitude estimation scales:** These scales focus rather on the difference between the individual test subjects than on an absolute value. Visual analog scales, for example, offer the rater a continuum to denote the respective property (cf. Figure 3.1 b). Starting from the first test subject the rater is asked to mark the other subjects in the continuum according to the relation

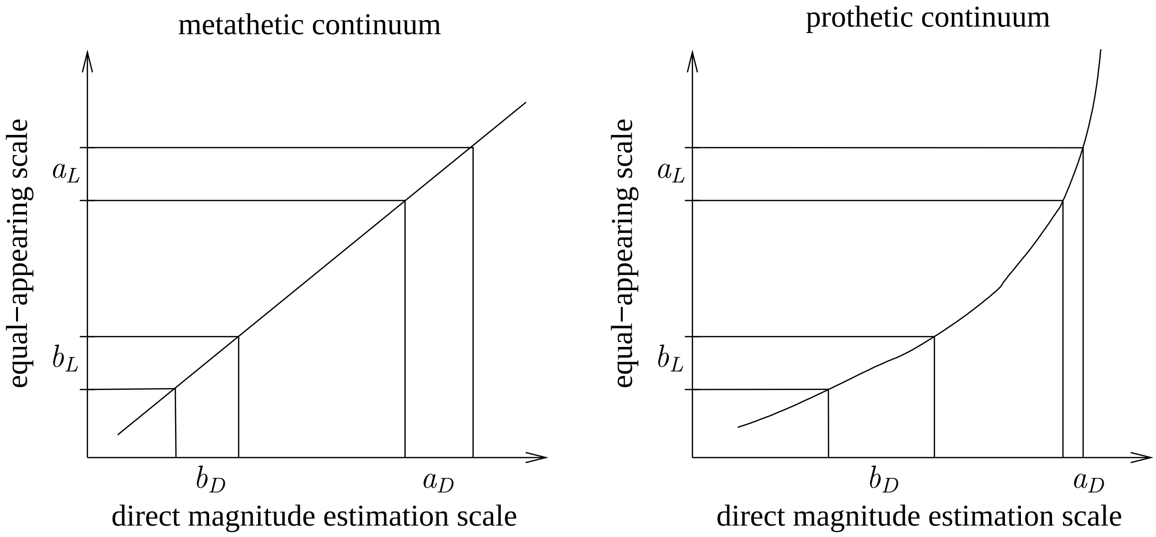


Figure 3.2: Difference between metathetic and prothetic continua: If the means of the ratings on an equal-appearing and a direct magnitude estimation scale are plotted against each other, the metathetic continuum shows a linear relation while the relation is typically skewed towards the end of the scale in a prothetic continuum [Steve 75].

between the current subject and the other subjects. So, a scale is obtained which represents the differences between the different test subjects.

According to [Steve 74] two different kinds of continua appear in subjective evaluations: Prothetic and metathetic ones. While a human being has no problems to partition a metathetic space into segments of equal size, a prothetic space cannot be partitioned by equidistant boundaries that easily. A typical example for a prothetic continuum is loudness while the pitch of a voice for example is a metathetic continuum. Pitch can be easily partitioned into equal-sized segments (from “very low” to “very high”) while it is difficult to assign such equidistant boundaries to the loudness. The question, however, which of two sounds is louder can be answered easily by a human being. Therefore, Stevens proposes to use direct magnitude scales if a prothetic continuum is to be rated.

A method to determine whether the continuum to be rated is prothetic or metathetic is to have two disjoint groups of raters label the same group of test subjects [Steve 75]. One group is given a direct magnitude estimation scale, while the other group performs the evaluation on an equal-appearing scale. Then the mean value is calculated for each test subject and the means are plotted against each other. In order to take the central tendency of the direct magnitude scale into account, the geometric mean is used while the arithmetic mean is taken for the equal-appearing scale. Figure 3.2 shows the effect: In the metathetic continuum, the relation between both scales is linear. In the prothetic continuum, the relation is skewed towards the end of the scale. So, in a prothetic continuum the raters fail to judge the data equidistantly using an equal-appearing scale. As can be seen in Figure 3.2, the equal-appearing intervals a_L and b_L are not of equal size. b_D is much larger than a_D in the prothetic case. However, this effect does not affect the rank of the test subjects as long as the curve is still monotonically increasing. The scores obtained on such a scale have to be interpreted rather as ranks than as absolute scores.

Due to this effect, the validity of equal-appearing scales for prothetic continua has been questioned in the literature [White 02a]. However, the demand that all research that has been done so far is wrong and has to be repeated with the proper scales is disproportionate. In fact, the use of direct magnitude estimation scales has also some disadvantages:

- It is difficult to find a certain point, e.g. the center on such a scale.
- Most raters are used to equal-appearing interval scales. In most rating scenarios, equal-appearing scales are used.
- Magnitudes obtained by this type of scale have to be interpreted rather as ranks than as absolute scores since the magnitude of two different raters is still different.

In order to avoid the problems of in-equal interval sizes, the perceptive scores can be mapped onto their ranks. Spearman's rank correlation [Spear 04], for example, also allows for a comparison in continua which are not equidistant.

In general, both scales still suffer from the same problem: The scale is highly dependent on the rater who performs the assessment.

3.1.1 Single vs. Multiple Labelers

In order to attenuate the high subjectivity of a single rater, often multiple labelers are asked to perform the same evaluation. So, the subjectivity of the individual rater is reduced by computation of the average of multiple raters. Sometimes, this procedure is already called "objective" in the literature. In the opinion of the author, the term *inter-rater-confirmed-subjective-mean-score* is more appropriate for this kind of score.

In clinical practice, the multi-rater evaluation is performed only seldom, because it is time- and manpower-consuming. Such an effort is often only done for scientific purposes. This leads to the following problems:

- The ratings have to be performed always by the same rater in order to be comparable.
- Ratings of a single rater still can show considerable variance.
- The ratings of two different labelers cannot be compared directly, because one rater might observe more strictly or is more lenient towards his patients.

So, an evaluation with multiple raters is always to be preferred to single labeler evaluations.

3.1.2 Expert vs. Naïve Labelers

Previous studies have shown that experience is an important factor that influences the perceptive estimation of speech disorders leading to inaccurate evaluation by persons with only few years of experience [Paal 05].

Naïve labelers can only rate data if they are given very clear instructions and if the task is simple. In intelligibility tests like [Zenne 86], these naïve listeners are asked to transliterate unknown words and phrases. Later on, the number of the correctly recognized words represents an intelligibility score. This procedure yields a score between 0 and 100 % and is often postulated to be much more reliable and appropriate for the measurement of intelligibility than equal-appearing interval and direct magnitude estimation scales [Schia 92, White 02a]. However, this type of evaluation has still several disadvantages:

- Each speaker gets to read different words and phrases. Therefore, the comparability between test subjects is limited because the distribution of the phones in the uttered data is not guaranteed to be the same. In fact, some phones might not appear at all.

- The naïve listeners get used to the limited number of words and phrases, or to the speech disorders of the patients. Thus, the rate of correctly identified words increases over time. After the assessment of a certain number of test subjects the constraint that the words and phrases are unknown to the listener is no longer met. Thus, either the vocabulary of the test has to be changed or the naïve listener has to be replaced.

- This often wrongly "objective" called procedure varies a lot between different raters: The mean values between different raters varied up to 21 percent points in our studies [Hader 07b] while the correlation, i.e., the agreement between the raters, was highly significant. For scientific purposes, also the mean of several raters is therefore required.

Parental questionnaires [Bosel 04] can also be found in the literature. These questionnaires are based on the idea that observations made by the parents allow for conclusions on the health status of their child. However, they are rather not sufficient [Fisch 05] because parents' evaluations are usually more influenced by emotional aspects and their experience.

The previously mentioned points lead us to the conclusion that naïve labelers only be used for very simple tasks. Again, the use of multiple naïve raters is to be preferred.

3.1.3 Aspects vs. Holistic Criteria

Experts in speech evaluation, like speech therapists, can judge many different details of the speech of a test subject. While holistic features, e.g. the speech intelligibility, have the advantage that they often can be quantified easily even by naïve listeners, aspects such as nasality have the advantage that they give a much better insight on the type and extent of the speech disorder.

An experienced speech therapist can distinguish all the different characteristics presented in Chapter 2. However, not all speech therapists can quantify these aspects. So, the evaluation of details often results only in qualitative labels instead of a quantification. Furthermore, as presented in [Sell 01], some of these aspects occur rather seldom. The expert might have to label very much data in order to find a representative number of the respective aspects of the speech disorder. Hence, the procedure might be extremely laborious.

The holistic labels can be created much faster, since the expert just has to state his impression on an equal-appearing interval or a direct magnitude estimation scale. The labeling of aspects is much more time-consuming and complex. Therefore, only selected characteristics of the speech disorders of children with CLP will be investigated in this work.

3.2 Objective Evaluation

For the objective evaluation of disordered speech, several methods can be found in the literature. Especially hoarseness has been investigated well in the literature. Algorithmic evaluation of the fundamental frequency started as early as 1902 [Buder 00]. Most of the algorithms, however, require special phonation protocols which might lead to misdetection of the hoarseness [Wurzb 06]. Nasality can be measured with devices like the Nasometer [Kay 94], or it can be detected in sustained vowels by analysis of the formant structure [Zevce 02]. Recently, the intelligibility of speech was quantized objectively. However, this was only done for adult speakers so far [Hader 04, Schus 05, Maier 07d]. In the following these approaches are discussed in detail.

3.2.1 Hoarseness

Hoarseness is a typical symptom for all kinds of voice disorders. Hence, it is suitable to describe the severity of a voice disorder. It is caused by an asynchronous oscillation of the vocal folds or insufficient closure of the glottis. Therefore, it can only be measured in voiced parts of the speech. Thus, most detection algorithms are based on the phonation of sustained vowels. According to [Buder 00], these algorithms can be divided into the following classes:

- Fundamental frequency (F_0) statistics
- Amplitude statistics
- Waveform perturbations
- Spectral measures
- Inverse filter measures
- Dynamics

Today, most of the algorithms used in scientific studies [Pahn 01] are implemented in commercial software [Kay 93]. While the evaluation of read or spontaneous speech still caused problems in the 1980's [Wendl 86], recent approaches can also handle this type of speech [Halbe 04]. According to [Wuyts 96], the examination of sustained vowels by many patients did not yield any correlations higher than 0.53 with perceptive evaluations of human experts. The reason of this effect might be dependency of the perception of the hoarseness on the frequency of the phonation [Wurzb 06]. The higher the frequency is the harder the hoarseness can be perceived.

The state-of-the-art evaluation method to measure the quality of a voice is the dysphonia severity index (DSI) [Wuyts 00]. In order to solve the problem of the frequency-dependency of hoarseness, the DSI states a multi-parameter approach. It is computed from several parameters which are measured during the phonation of vowels. Then, maxima and minima of these parameters are combined to represent the static as well as the dynamic features of a voice.

Using the linear discriminant analysis (LDA, cf. Chapter 4.2.3), Wuyts selected the best subset of 20 features computed with commercial software [Kay 93] to create the DSI. In order to get classes for the LDA, the data was partitioned perceptively into 4 groups according to the G (hoarseness) of the GRBAS scale [Hiran 81] by speech pathologists. The DSI was obtained from the most discriminating component of the LDA. In total, four parameters were selected to compute the DSI:

- The **maximum phonation time** (t_{MPT}) taken in seconds is an aerodynamic parameter and is measured while the patient is asked to sustain a vowel as long as possible. It can be interpreted as the ability of the subject to maintain continuous speech. While control subjects have a t_{MPT} of 18.9 ± 6.7 s in average pathologic speakers sustain vowels only for 12.4 ± 6.4 s in average.

- The **highest fundamental frequency** the patient is able to produce ($F_{0\ \max}$ in Hertz) gives an upper boundary of the patient's voice. A patient with dysphonia can produce only vowels in a limited range of the F_0 compared to control subjects.

- The **lowest sound intensity** (I_l) which is emitted during the phonation by the patient in dB(A) can be interpreted as the "gentleness" of the voice. If the vocal folds are obstructed in any way, much more energy is needed to make them oscillate. Therefore, the patients have difficulties to pronounce soft speech while normal speakers have no problems in speaking with a low intensity.

- The **jitter** (J) of a voice denotes the mean deviation of the period length and is measured in % of the mean period length. Although there are many definitions of jitter Wuyts et al. do not state which of them they are using. In this work the jitter is extracted as described in [Levit 00, p.14]. Jitter is perceived as roughness in the sound of the voice. In a normal voice, the jitter is 0.7 ± 0.5 % in average. In pathologic voices, however, the jitter is 2.6 ± 2.3 % in average [Wuyts 00].

With these four parameters, the DSI is now computed as:

$$\text{DSI} = 0.13 \cdot t_{\text{MPT}} + 0.0053 \cdot F_{0\ \max} - 0.26 \cdot I_l - 1.18 \cdot J + 12.4 \tag{3.1}$$

A DSI of 5 is regarded as an excellent voice while -5 is the result of a very dysphonic voice. Values outside these boundaries might occur as well in extraordinary voices. A DSI ≥ 1 is regarded as normal. If the fundamental frequency cannot be measured—like in patients without larynx—the DSI is considered to be -5.

The correlation of the DSI to the G of the GRBAS scale is very high (0.996). Unfortunately, the article does not clarify whether the training data was different than the test data which seems not to be the case. In order to compensate this flaw, the authors compare their DSI to another perceptive evaluation method: the voice handicap index (VHI, [Jacob 97]) which states the severity of the disorder in the view of the patient. So, the authors obtain another significant correlation of -0.79 ($p<0.001$) between the DSI and the VHI which only proves that there is a high correlation between the G of the GRBAS scale and the VHI in the data they used. However, this does not confirm the validity of their index.

Nevertheless, the DSI is the most common and widely used "objective" index in the literature for the evaluation of hoarseness. New and better methods for the objective measurement of the hoarseness of patients are being developed [Dolli 02, Dolli 08] which will allow the investigation of non-stationary phonation [Rasp 06] and visualization using for example the phonovibrogram [Lohsc 09, Eysho 08].

3.2.2 Nasality

The term nasality is often used in the literature for two different kinds of nasality: hyper- and hyponasality. While hypernasality is caused by enhanced nasal emissions, like found in CLP children (cf. Chapter 2), hyponasality is caused by a blockade of the nasal airway, e.g. when a patient has a cold. In the literature studies on both effects are found [Hardi 92]. However, most of them concern only the effects on voiced speech (vowels) [Pruth 03, Pruth 04, Pruth 07a, Pruth 07b, Pruth 07c] and consonant-vowel combinations [Cairn 96a, Katao 96].

Figure 3.3 shows the effect of the vowel nasalization in the model spectrum [Atal 67, Atal 79] of the vowel /a:/. In both spectra a slight nasal formant $F_1^N(\omega)$ exists between 300 and 500 Hz. The first formant $F_1(\omega)$ is at about 1100 to 1300 Hz. In the nasal /a:/, the intensity of the $F_1^N(\omega)$ is stronger than the $F_1(\omega)$ which causes the nasality to be audible. Actually, this effect is caused by a combination of the following effects [Fant 60b]:

- The first formant bandwidth increases while the intensity decreases.
- The nasal formant $F_1^N(\omega)$ appears or is increased.
- Antiresonances appear which increase the strength of the antiformants $F_k^A(\omega)$.

According to [Cairn 96b] these effects on the energy of the normal speech (S) and hypernasal speech (S_{nasal}) can be modeled by a combination of formants at various frequencies:

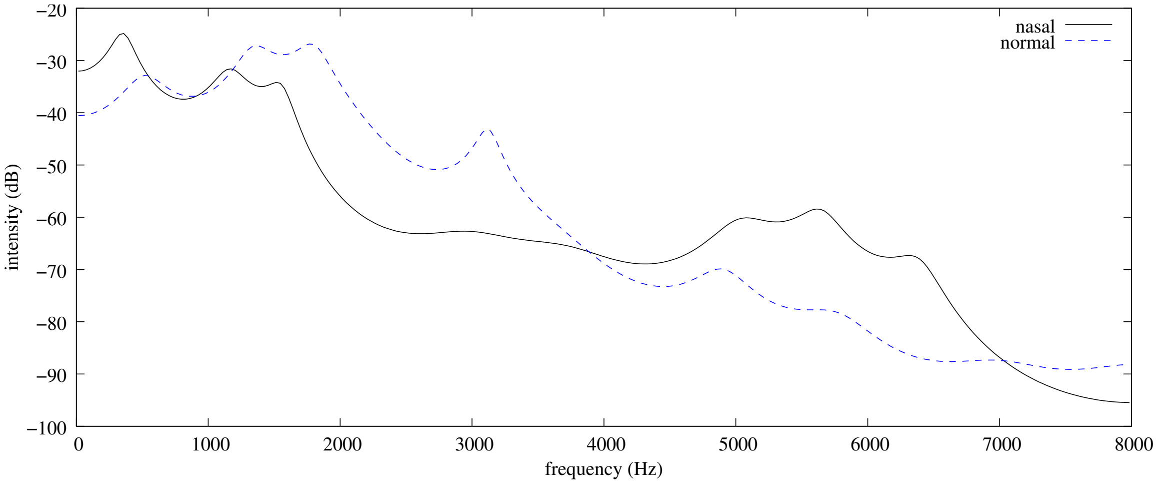


Figure 3.3: Model spectrum of a nasal and a non-nasal realization of the phoneme /a:/ in the phonetic context /ha:s@/ using 20 coefficients: The nasal formant F_1^N (∼300 – 500 Hz) is stronger than the first formant F_1 (∼1100 – 1300 Hz) in the nasal realization. Note that the displayed speech is children's speech which causes exceptionally high formant frequencies.

$$S(\omega) = \sum_{i=1}^{I} F_i(\omega) \tag{3.2}$$

$$S_{\text{nasal}}(\omega) = \sum_{i=1}^{I} F_i(\omega) - \sum_{k=1}^{K} F_k^A(\omega) + \sum_{m=1}^{M} F_m^N(\omega) \tag{3.3}$$

Here, $F_i(\omega)$ denotes the intensity of the i^{th} formant in the frequency domain. According to the literature, the main cause for nasality is the intensity reduction of the first format [Fant 60b, Zevce 02]. Using low-pass and bandpass filters to compute a hypernasality feature as described in Chapter 4.2.4, Cairns et al. achieved classification rates of 94.7 % for normal and hypernasal speech in vowel /i/ and 93.0 % for normal and 93.3 % for /A/ [Cairn 96a].

In [Haapa 96] features computed from a model spectrum with 14 components are used to classify realizations of the vowels /i/, /u/, and /a/ to either hyponasal or normal. The classification system showed good agreement with the perceptual evaluation.

Zečević introduced a framework for the automatic evaluation of the nasality in children [Zevce 02]. Unfortunately, the evaluation environment consists of three components which were never integrated into one system. The formant tracking algorithms were all implemented in MatLab while recording was performed using MacOS. The use of speech recognition techniques was intended but never applied since the author had only access to commercial dictation software which yielded a bad recognition on the children's speech data. All that was done was classification of nasality to the classes "normal" and "hypernasal" speech according to the position, intensity, and bandwidth of the formants, antiformants, and nasal formants on sustained vowels. Classification rates of 70 to 80 % were achieved.

Analyses on continuous speech started as early as 1964. However, all these methods are invasive to some extent. Warren et al. inserted two catheters connected to a differential pressure transducer into the nose of the patient [Warre 64]. Thus, they could measure the nasal air flow below and above the velopharyngeal orifice. From the difference in the pressure flow, an estimate value of the area of the velopharyngeal orifice can be calculated. It is shown that this estimate varies less than 10 % from the actual size of the area.

In order to calculate the Horii Oral Nasal Coupling (HONC) index [Horii 81], an accelerometer which measures the vibration of the tissue has to be attached to the external surface of the nose. A second accelerometer is attached to the throat of the patient. So, the fraction of the nasal air emission can be calculated. Since this fraction is highly dependent of the setup and the patient, the system has to be calibrated before the measure. Therefore, the patient has to phonate /m/ in order to get a sound with a maximal nasal emission. All following measurements are computed with respect to this maximal value. Good consistency with the perceptual evaluation of expert raters could be shown in [Horii 83]. Further modifications of this technique lead to the Nasal Accelerometric Vibrational Index (NAVI) [Reden 85] and the Nasal Oral RAtio Meter (NORAM) [Karli 93a].

In clinical practice the most wide-spread objective method is the measurement of the nasal air flow using the nasometer [Kay 94]. Based on the oral-nasal measures of Fletcher et al. [Fletc 89], the nasometer consists of two microphones which are separated by a metal plate. One of the microphones is placed in front of the nose and the other one in front of the mouth. The metal plate is used to mutually insulate the microphones. The so called nasalance denotes the fraction of the nasal air flow divided by the oral air flow. High correlations are reported in the literature between the nasalance and the perceptive evaluations by human experts [Hardi 92]. Furthermore, it is possible to partition normal speech from pathologic speech with the nasometer [Stell 94, Karne 95]. The nasalance measurement was shown to be applicable to other languages than English as well [Tachi 00, Bress 99a, Kuttn 03].

Recently, other measurements were inspected as well. Using videopharyngoscopy [Ysunz 03] the area of the velopharyngeal orifice can be inspected directly without having to measure different air flows. However, the method is invasive again, and it is not tolerated by every child, especially not by young children.

A very interesting approach is described in [Lu 04]. The author uses Gaussian Mixture Models (cf. Chapter 4.2.4) and support vector machines (cf. Chapter 4.2.2) for the classification of continuous speech into the classes “normal” and “hypernasal”. However, the classification rate of 94.85 % seems very high for this difficult problem. A closer look on the experiments reveals that in total three databases were collected for the work: two containing non-nasal and one containing nasal speech. For the experiments one of the non-nasal databases and the nasal database was split into a training and a test set. Therefore, neither the spoken sentences nor the language were matched (one non-nasal database contained welsh language). Whether the same microphone was used for the three databases is unclear. Presumably the high classification rates can be ascribed to the large differences in nasal and non-nasal databases. But it is also possible that the classifier at least to some extent learned the channel than the “nasality” of the speech data.

3.2.3 Intelligibility

In the literature transcription tasks such as the Post-Laryngectomy-Telephone-Test (PLTT) are often called “objective” [Zenne 86]. In the test the test subjects are given a random set of words and phrases which they have to read. Then a human “naïve” listener who does not know the words notes the words down he understood. The difference of the target words and the understood words forms a score which represents the intelligibility. In the author’s opinion this is not the case. Experiments of Haderlein et al. [Hader 07b] showed that transcription tasks as the PLTT can vary a lot between different transliterating persons. Hence, the inter-rater reliability is also restricted as in experts’ evaluations.

Therefore the only objective method in terms of the definition above found in the literature is the one of Haderlein et al. [Hader 04, Hader 07a]. In order to simulate a naïve listener the authors apply a speech recognition system (cf. Chapter 4.2.4). To evaluate the method, the recognition rate of the word recognizer was compared to perceptive evaluations of expert listeners [Schus 05]. The method was tested on speech of patients after removal of the larynx (laryngectomees) whose spectral characteristics differ a lot from normal speech [Robbi 86]. The authors could show a high correlation between the experts’ opinion and the word accuracy of -0.84 for a group of 18 speakers [Schus 06a] and -0.88 for a group of 41 laryngectomees [Riedh 06].

An automatic version of the PLTT is presented in [Riedh 07a]. The perceptive evaluation showed a high inter-rater variability in the absolute scores of the test. A speech recognition system was employed to model an additional naïve listener. This time a telephone speech recognizer was employed [Riedh 06]. The outcome of the speech recognizer was compared to human evaluations of the PLTT. The automatic recognition system as well as the naïve individuals showed a high correlation with the average of the naïve listeners’ scores.

The technique described above also showed very good consistency with experts’ scores in other patient groups. The measurement of the intelligibility could also be applied to patients with articulation disorders due to cancer in the oral cavity [Maier 07d].

The same group showed that prosodic features (cf. Chapter 4.2.4) extracted from the speech data of the laryngectomees also hold a lot of information about the speaker [Hader 06a]. They could show that, for example, the effort in speaking and the correctness of the breath-sense units correlates with the length of the pauses after a word and the onset position of the fundamental frequency. In [Maier 07a] it was shown that prosodic features and the recognition rate of a speech recognizer can be successfully combined to improve the prediction of intelligibility scores given by speech therapists.

The Sammon mapping is a technique for non-linear dimension reduction (cf. Chapter 4.2.3). It computes a low-dimensional point cloud which preserves the topology of the high-dimensional space. However, since speech varies a lot in time a representation of the speech data with a fixed number of parameters has to be found. Therefore, speaker adaptation techniques are applied. The idea is that a speaker-adapted speech recognizer not only improves the recognition for a single person but also represents the speaker himself. Since speech recognizers have a lot of parameters, their mutual dependency is difficult to understand. The Sammon mapping can be applied to find a visualization of the dependencies between the recognizers in a low-dimensional space [Shoza 04]. In [Hader 06b] it was shown that this representation displays properties of the pathology of a person's speech as well. The resulting map separates normal from pathologic speakers. Furthermore, it is shown that the position of a person on the map allows conclusions on the grade of the intelligibility to some extent.

3.3 Towards a Ground Truth

In the previous section, we discussed the pros and cons of objective and perceptive evaluation methods. The major pro of the perceptive methods was that they can be applied easily and they are non-invasive. However, all perceptive measures lack objectivity. Even when the average of multiple raters is taken into account the measure is still subjective and rather not reproducible. Furthermore, detailed evaluations have to be done by experienced raters. Attempts to omit this problem, such as transcription tasks, cause other problems, e.g. naïve raters have to be replaced after they got used to the text and the disorder to keep the constraint of naïvity, and thus none of the naïve ratings is reproducible. Additionally, the transcription tasks have the same inter-rater differences as all other perceptive evaluations.

Objective evaluations have the advantage that they are reproducible, reliable, and fast in most cases. However, one has to keep in mind that some tests are rather unspecific, like the evaluations of sustained vowels, and others are invasive which might be uncomfortable for the patient. A solution to this problem is the use of automatic speech-based classification systems. However, these systems have to be trained with a so-called "gold standard" or "ground truth" which are virtually always derived from perceptive evaluations.

In order to get a valid gold standard for an automatic system, one should always pick as many raters as possible and the correct procedure according to the type of the continuum to be rated. However, one has to keep practical considerations in mind as well. The creation of detailed labels is often a very laborious process for an expert rater. If the ratings cause a lot of effort to the rater, it is difficult to find many raters for this aspect. Therefore, very detailed qualitative labels can be done only for a small database since the effort to the labeler is very high. Since all perceptive evaluation methods are subjective and tampered by inter-individual differences, the best trade-off between accuracy and effort has to be chosen. Quantitative evaluations have to be interpreted as rank anyway since the absolute value obtained by computation of the average of multiple raters is irreproducible and therefore questionable. Thus, correlations of the ranks should rather be investigated than correlations of the absolute values (cf. Chapter 4.2.1). Many different aspects might spoil the validity of a gold standard. Nasality, for example, is a prothetic continuum if it is rated by human raters [White 02b] and should be evaluated rather qualitatively than quantitatively. Furthermore, hoarseness might tamper the perceptual evaluation of nasality [Imato 99]. Therefore, nasality can only be evaluated in a non-hoarse voice. If these effects are regarded properly and the scale and rating procedure are chosen carefully, a good gold standard can be obtained and used for the training of an automatic evaluation system.

Chapter 4

A Program for the Evaluation of All Kinds of Speech Disorders

At the beginning of this chapter, an overview on the Program for the Evaluation of All Kinds of Speech Disorders (PEAKS) is given. It describes the architecture, the components, and of course the security and collaboration concepts needed in a clinical environment.

Next, the algorithmic background and the mathematical concepts which are being used throughout this work have to be clarified. The section will describe different measurements of agreement, support vector machines, dimension reduction techniques, speech processing, and the normalization of age effects.

After having discussed the architecture of the system and the algorithmic background, the architecture of PEAKS and its implementation details will be presented at the end of this chapter.

4.1 System Overview

In order to develop a system for the automatic assessment of children's speech, several points have to be considered. These can be formulated as requirements which should be implemented in the final system.

Starting from these requirements, the abstract use cases of the system are formulated (cf. Chapter 4.1.2): patient and user management.

The system's setup portrays the client-server architecture of the PEAKS system. Furthermore, the configuration of the components and the integration of the requirements are described.

Security and collaboration concepts are vital for a system which is to be used in a clinical context. They are presented in Chapter 4.1.4.

4.1.1 Requirements for a Speech and Voice Evaluation Platform

A system which is going to be used in a clinical environment to analyze speech disorders of children has to fulfill certain requirements. First of all, the system must be able to perform different analyses with the speech data. In our case the system should enable the following analyses:

- **Intelligibility Measurement:** The system should be able to compute an estimate intelligibility score. This score should correlate highly to the scores of a human expert.
- **Pronunciation Assessment:** The assessment of the pronunciation, i.e., articulation disorders is another crucial point in the list of requirements of a speech evaluation platform. Again, the system should deliver an analysis which is in accordance to a speech therapist's opinion.
- **Visualization:** The visualization of the dependencies between different speakers gives a better understanding of the speech disorder and helps the medical personal to compare different graduations of speech disorders with each other.

Besides these main points, a medical evaluation system should implement further passive requirements in order to be applicable in a medical context:

- **Multi-user Support:** A system to be used in a clinic should provide access to different users simultaneously. Multiple examinations can be performed at the same time.
- **Platform Independency:** Platform-independency will allow the system to be used on multiple operating systems. This is necessary because the computer networks of a clinic are often heterogeneous and use more than one operating system.
- **Security Concepts:** The patients' data are confidential and must only be accessible to the treating clinician. Security leaks have to be avoided. Therefore, encryption features should be integrated into the platform.
- **Collaboration Concepts:** A platform for the evaluation of speech disorders should provide methods to allow for the collaboration of multiple users. Multiple clinicians can provide treatment to the same group of patients. Furthermore, the system should implement interfaces to perform an export of the data in order to be used by other programs. For example, statistic analyses for a scientific study can be performed with the software the user is used to.
- **User Comfort:** The system should be available to the clinicians' PCs easily. Complicated installation procedures and software incompatibilities cause unnecessary trouble. Therefore, the software should be as easy to use as possible.

4.1.2 Use Cases

Use cases are a wide-spread technique to develop software and document the functional requirements needed by the different actors, i.e., users of the system [Rupp 06]. Usually, a use case reports the actors who are concerned by the action, the preconditions, a short description of the action, and the resulting postconditions. These

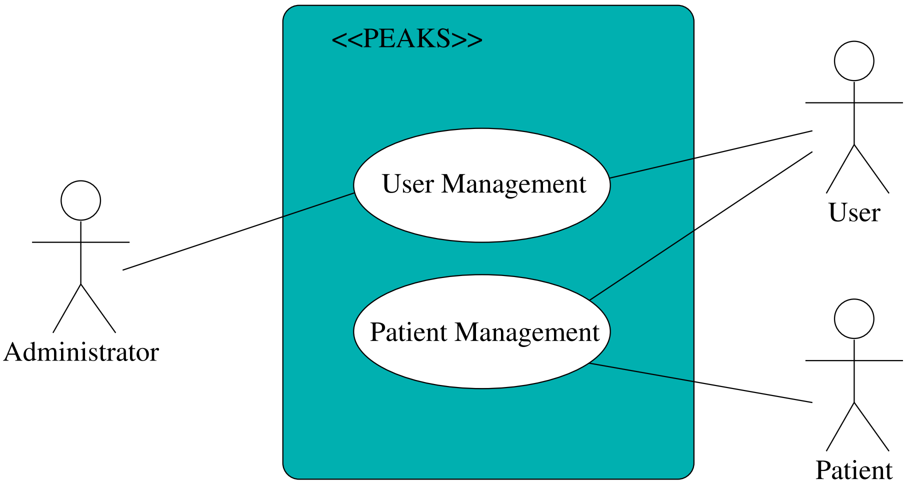


Figure 4.1: Use case diagram of the PEAKS system: The diagram shows the two main functions which are available to the different actors: user management and patient management.

are especially useful if the development and the implementation are done by different engineers. Since this is not the case for this work the actual use cases were not formulated in detail. However, a very practical technique to summarize the use cases are use case diagrams. A use case diagram displays the system, the actors, and the possible actions i.e. use cases of the system. Figure 4.1 shows the different abstract actions which can be done by the users according to their roles within the PEAKS system. The diagram shows two abstract use cases: patient management and user management.

- **Patient Management:** Most of the actions which can be performed within the PEAKS system are designed to handle patients. A user can evaluate recordings of a patient, run the automatic assessment routines, and create reports and visualizations. The patient has just a passive role within the PEAKS system. New patients can be created and recorded by a user of the system. This is the actual use case of the system and is described in Chapter 4.1.3.

- **User Management:** The users themselves are responsible for the registration to the system and their authentication. However, an administrator is still needed to keep the PEAKS service running. All security-relevant tasks have to be conducted by him. He is responsible for the setting of permissions and the creation of new passwords, in case a user forgot or lost it. This is all part of the security functions which are described in Chapter 4.1.4.

4.1.3 System Setup

PEAKS can be interpreted as a pattern recognition system after [Niema 03] which is divided into a client and a server side as can be seen in Figure 4.2. The client software runs on any standard PC with Internet connection. It is used to manage the different patients, create new recordings, to review the results of the different analyses, and to

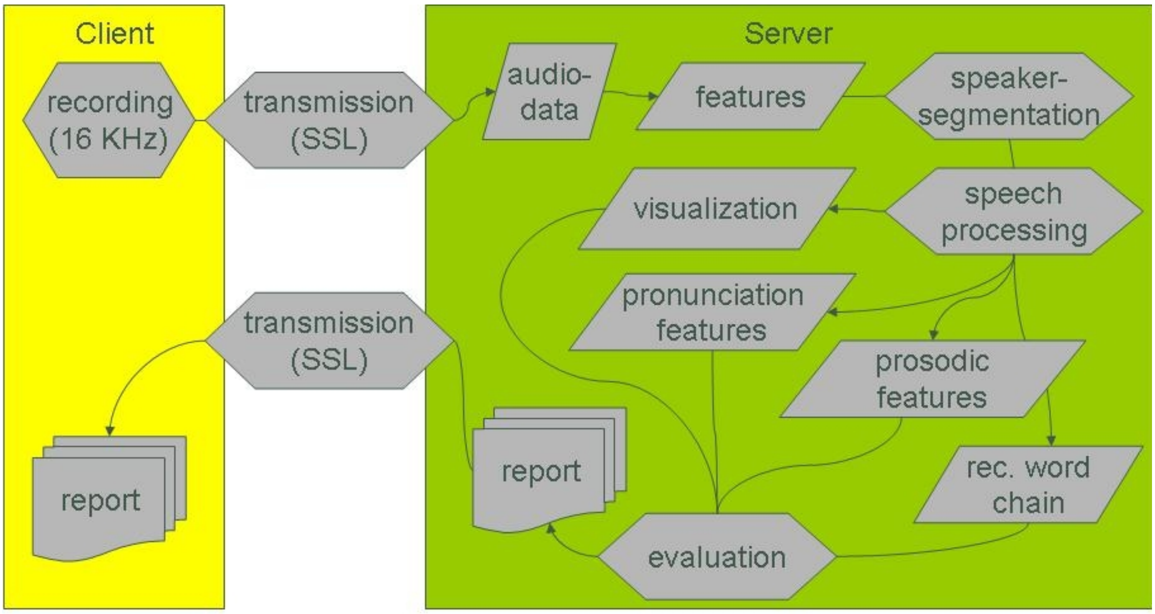


Figure 4.2: The Architecture of the PEAKS system: While access to the data is provided by a thin client all the performance intensive speech analyses are computed on the server.

visualize the dependencies between the different patients. Because it is designed as an applet, it runs in any web browser with Java^{TM} support.

All computationally expensive tasks are performed on the server: The preprocessing, the feature extraction, the assessment, and the computations necessary for the report generation. Only the result of the analyses are transmitted to the client.

This setup has several advantages compared to a stand-alone application:

- **Software Updates:** New versions of the software are downloaded at the start of the applet automatically. Therefore, all client PCs always have access to the newest software without the user even noticing.

- **Low Hardware Requirements:** Java^{TM} applets have only minimal requirements to the hardware of the client PC: The maximal memory space occupied by an applet is 64 MB by default.

- **Platform Independency:** Java^{TM} applets provide multi-platform support. Therefore, the system can also be applied in a heterogeneous computer network with multiple different operating systems.

- **Software Reuseability:** Most of the speech processing software of the Chair of Pattern Recognition is written in C for Linux operating systems. Since different operating systems can be used on client and server, all the Linux software can be installed on the server system which is not visible to the client side.

However, the client-server setup has also some disadvantages which have to be compensated:

- **Parallel Data Processing:** Data can be received from multiple clients simultaneously. Therefore, one has to pay special attention to process and data management in order to avoid deadlocks and data inconsistencies.

- **Dependency on the Internet:** An Internet connection is required to use the system at all times because the data have to be transferred to the server in order to be processed.

Process and Data Management

All tasks which are performed on the client system must be able to be performed at the same time. For each connecting client, a new thread is spawned which handles the client's requests. Parallel requests are no problem so far since the back-end database can handle simultaneous requests and the audio data are stored in the server's filesystem using different names for each patient and recording.

However, most of the speech processing software of the Chair of Pattern Recognition was never designed to work concurrently. In order to compensate this effect, the server keeps track of the analyses which were requested by the different clients in a job list. Then the analyses are processed in a first-in first-out manner. Another side effect of this procedure is that there is always just a single process with high processor load on the server's CPU. There is still performance left to run the PEAKS server's tasks and the web server even on a machine with just two processors.

In order to differentiate between the users, they have to register to the system by choosing a username and a password. This information identifies every user of the system uniquely.

Data Acquisition

Data acquisition is performed on the client side. In our scenario certain pictures or words are presented to the subject and their speech is recorded. In the following this procedure is called a "test". To perform a test, a standard PC with a sound card and microphone is sufficient. However, it always has to be connected to the Internet. Then the user logs into the system, selects a patient and a test and performs the recording. The recorded data is then sent directly to the server.

Unfortunately, many places where the data acquisition took place could not provide Internet access. Therefore, an offline version of PEAKS was developed: PEAKS-local. As Figure 4.3 shows, it simulates a full PEAKS server for a single user. The following three steps have to be performed to do recordings locally:

1. **Initialization of the Server:** The Server has to be initialized in order to allow local recordings over the Internet. During the initialization procedure, all tests which can be performed by the system are downloaded to the local file system. Furthermore, the information about the patients the user has access to are downloaded. This enables the user to create another recording of an already existing patient. For this reason the user has to input his username and password during the initialization process.

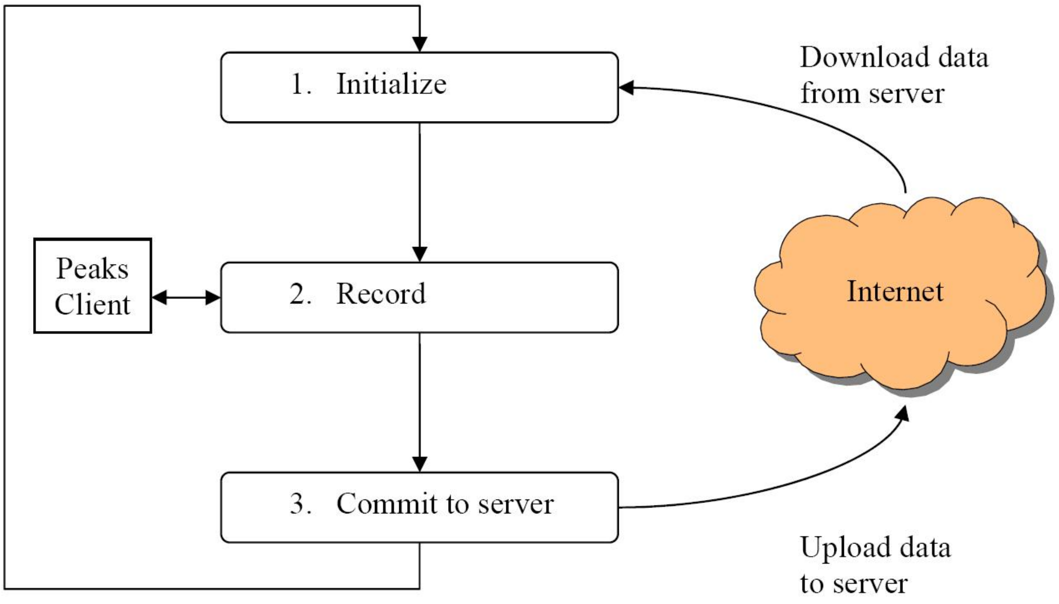


Figure 4.3: The Architecture of the PEAKSlocal system: The normal PEAKS client is connected to a simulated PEAKS server which has just the ability to store new recordings. Later on, the recorded data can be committed to the real PEAKS server.

2. **Local Recordings:** After the server was initialized, the user can create local recordings. Instead of the real PEAKS server, he connects to the PEAKSlocal server. Since the PEAKSlocal server supports only a single user, username and password are not required to log into the PEAKSlocal server. Unlike the PEAKS server, the local server supports only the creation of new users and the recording of new tests.

3. **Commit of the Server:** After the recordings took place, the data have to be committed to the PEAKS server. Therefore, Internet access is required again. During the commit procedure, the local server transmits all of the newly acquired data to the server. Lastly, the local audio and user data is deleted and the local server remains uninitialized.

Another problem in the data acquisition are the different acoustic conditions in which the recordings take place. For example, noisy fans inside the recording computers and cheap microphones are to be avoided. Very expensive microphones and sound hardware, however, pose the problem that they are often difficult to use. Furthermore, microphones using a microphone stand collect a lot of the sound events in the environment which can alter the result of the analyses.

For our purpose, headsets work best for the acquisition of the data. On a standard PC, a normal headset was used. Data acquisition using mobile computers often posed the problem that their power adapters could not filter out the frequency of the AC power during AC/DC conversion properly. These frequencies were still audible on the recorded sound track. The use of a USB headset with integrated analog/digital conversion solved this problem on most mobile PCs.

Preprocessing

Preprocessing is the second step in a classification system (cf. [Niema 03, p.26]) directly after the acquisition of the data. Its object is to improve the signal quality before the feature extraction step. Unlike feature extraction which results in a feature vector and represents the most important aspects of the signal, the preprocessing is intended to preserve most information of the signal, i.e., it is still suitable to be presented to humans.

In intelligibility assessment the preprocessing is integrated into the feature extraction step and can hardly be separated from it. The step during feature extraction which could be part of a preprocessing as well is the noise reduction using cepstral mean subtraction. Since linear distortions are additive in the cepstrum of the signal, they can be removed by subtraction. The basic idea is to model the distortion as the mean value of the cepstrum of a certain number of time steps. Then for each time step the distortion is subtracted from the feature vector [Schuk 95].

For the assessment of single words, quite a lot of preprocessing has to be done: First the words which are relevant for the test have to be spotted and segmented. Since the children tend to use *carrier sentences* like "This is a . . . " which are not part of the test. These must be excluded from the further processing. Then, the speaker has to be identified since the young participants often have to be encouraged by the supervisor during the test. In virtually all tests the supervisor's voice is audible. In order not to accidentally assess the words of the supervisor, his words have to be removed from the data. Chapter 4.2 describes these necessary computations in detail.

Feature Extraction

As features to represent the properties of the speech data we apply many different algorithms. While we give just a brief overview here, the exact computation of these features is presented in Chapter 4.2.

- **Mel Frequency Cepstrum Coefficients (MFCC):** The most commonly used features in speech recognition [Stemm 05, Gallw 02, Stemm 01, Schuk 95] which have already been successfully applied to speaker and speaker group identification [Bockl 07c, Yang 04, Reyno 95], detection of hypernasality [Maier 07c], and the recognition of emotions [Schul 07].

- **Prosodic Features:** Primarily, prosodic features were designed to model the segmental and suprasegmental speaking style [Noth 91, Niema 93]. Therefore, the fundamental frequency, energy, pauses, durations, jitter, and shimmer are extracted from the signal. Using these features it is possible to detect the boundaries between phrases [Batli 95]. Furthermore, prosody can be used to detect emotions [Huber 02], the user state [Adelh 03, Batli 03b], and the focus of attention [Hacke 06]. Even in pronunciation scoring [Hacke 07a] and speaker recognition [Chen 06, Chen 05], prosodic information could improve the recognition further.

- **Pronunciation Features:** Further information about the pronunciation is modeled using pronunciation features. In most scenarios they are used in computer-assisted pronunciation training (CAPT) or computer-assisted language learning (CALL) [Hessl 05, Hacke 07b]. The features as described in [Hacke 05b] are independent of the first language of the language learner. This is done by the computation of phone confusion tables, likelihood scores, duration features, and recognition accuracy. Another approach to model pronunciation variations of a language learner is to model often occurring pronunciation errors. This, however, has to be done for each first and second language pair individually. An example for German and English is given in [Hessl 05]. A thorough description of pronunciation and pronunciation variants of different languages is found in [Canep 05].

- **Teager Energy Profiles (TEP):** The Teager operator [Teage 90] is used to model non-linear events in speech and is highly sensitive to multi-component signals. As shown in [Cairn 94], it can be used to detect stress in speech data. Furthermore, in [Cairn 96a] the proof was given that it can also represent nasal events in speech. Again the Teager Energy Operator's sensitivity can be exploited since nasal speech below the second formant is a multi-component signal while normal speech is not (cf. Chapter 4.2.4).

- **Word Recognition Rate and Word Accuracy:** The word recognition rate and the word accuracy are also used as features in the PEAKS system. Actually these rates were designed to measure the performance of a speech recognition system [Schuk 95]. In our case we turn their function upside down: Since the recognition performance in normal speakers of our recognition system was validated before, we can estimate the intelligibility from the number of misunderstood, i.e., not recognized words. As previously shown [Hader 04, Schus 05, Schus 06a, Maier 07d], the word recognition rate and the word accuracy have significant and high correlations with the mean opinion of multiple experts.

Assessment

From the different features, the scores for each child can be computed. These can be obtained either by classification or regression analysis. Classification assigns the child, represented as a vector of features, one of K classes. In PEAKS the following methods are implemented:

- **Gaussian Mixture Models:** These models are standard classifiers in pattern recognition. The idea is to model the distribution of the feature vectors as a combination of a fixed number M of Gaussian densities. The mean values and the covariance matrices are estimated using the Expectation-Maximization (EM) algorithm [Demps 77].

- **Support Vector Machines:** The basic idea of Support Vector Machines is to model the border between two classes by the feature vectors which are closest to the class border. These vectors are called Support Vectors. Only the Support Vectors have to be saved instead of the whole training set. Classification is then performed by selection of the closest Support Vector. During this procedure different kernels can be applied in order to model non-linearities in the data [Schol 97].

In the case of regression analysis, the score is computed as a best estimate from the training data. The assigned value, however, is not a class but a floating point value. For this kind of prediction, the following methods are integrated into PEAKS:

- **Multiple Regression Analysis:** Since Pearson's regression [Pears 96] is only defined for 1-D inputs and 1-D outputs, i.e. target values, the method is not applicable to multivariate data. In [Cohen 83] an extension to multidimensional inputs is formulated. Therefore, a parameter vector is determined which produces the best linear estimate of the 1-D outputs by the computation of the Moore-Penrose pseudo-inverse [Moore 20, Penro 55] of the data matrix. Using this technique, a multiple correlation analysis can be performed as well by calculation of the correlation between the estimates and target values.

- **Support Vector Regression:** In a similar method to Support Vector Machines, Support Vectors can also be used to determine a set of parameters in order to predict a target value from a given input. Unlike the Multiple Regression Analysis, the Support Vector Regression uses the outliers to compute the regression weights. Therefore, this kind of regression is known to be more robust to outliers.

Report Generation

A major aim of the PEAKS system is the creation of reports which give an easy access to the information. Therefore, a report is generated which has several features:

- **Intelligibility Score:** The intelligibility is presented as word accuracy and word recognition rate. Furthermore, both are displayed using a marker in the respective value's distribution of the patients of the same patient group.

- **Phonematic Assessment:** An evaluation of the patient's speech is computed for certain articulation disorders and reported as a number of detected events. The kind of pronunciation score can be selected from a panel of different pronunciation assessment tools which are reported in more detail in Chapter 4.3.

- **Visualization:** The speaker dependencies are visualized on the report using the Sammon Mapping (cf. Chapter 4.2.3). It displays the parameters of a speaker-adapted word recognizer which represents the speaker. In order to reduce the dimensionality, the parameters are mapped to 2-D or 3-D points which preserve the distances of an arbitrary measure best. As shown in [Hader 06b], this method produces maps which have sensible axes representing e.g. the intelligibility.

- **List of Affected Phones:** At the end of the report, a list of the phones which were confused the most by the recognizer is found. It can be used to draw conclusions on the type and extent of the speech disorder.

4.1.4 Security and Collaboration Concepts

In a medical environment, the security of the data is a crucial point. The patients' privacy must be kept under all circumstances. Therefore, PEAKS implements a couple of security methods presented in the following.

Furthermore, PEAKS provides access permissions on user level. Users can register to the system and can start recording instantaneously. Later on, the access to a user's patient can be shared with other users in order to enable collaboration between different users.

General Security Concepts

PEAKS has several built-in security features which ensure the safety of the patients' data:

- **Secure Transmission:** All data which is sent between the client and the server is encrypted using the Secure Sockets Layer (SSL, [Resco 01]). SSL is a hybrid encryption protocol. SSL is located above the transport layer in the ISO OSI layer model [Zimme 80]. It works transparently. The encryption is not visible to the application layer. Hence, it is easy to use.

  The end-to-end connection uses a symmetric key algorithm in order to encrypt the data. The message integrity is kept by computation and transmission of check sums.

  Asymmetric encryption methods are also optionally provided for the authentication of the hosts during the establishment of the connection. This is usually done with a hand shake.

- **Timeouts:** In order to keep the patients' data safe in a clinical environment, the connection to the client system times out after a period of 20 Minutes of no communication with the client.

  This is sensible since in a medical environment emergency situations might appear in which the user forgets to log off properly from the system. This could be a security hole since an intruder might take advantage of the emergency situation and take control of the abandoned client system.

- **Pseudonymization:** The patient's name and other sensible information must not be entered into the system. Instead, a pseudonym is used to identify the patients.

  Common methods to create such an alias is to use either the patient's ID obtained from the patient administration software of the clinic or user hashes[1].

[1] Such hashes can be created by any combination of a fixed number of letters from the patient's first and last name and a fixed number of digits from the patient's date of birth. For example, using the first letter of the first name and the last name plus six digits of the date of birth yields "AM261180" for the author of this work.

Access Control

The access to the data should be limited for all users of the system in order to keep the data of each single user safe. Therefore, each user has to register with the system. Furthermore, it should be possible to allow certain users to share the data they are working on.

- **Registration and Identification Process:** In order to register with the system, the user has to fill out a registration form via Internet. The form asks for an unique username, a password, an e-mail address, a standard working directory, and the user's date of birth. Right after the completion of the form the user's speech is recorded. This enables the system later on to automatically exclude the user's speech data from the speech evaluation process with speaker identification techniques.

  All data is then transferred to the server. The password is transformed to an MD5-hash [Rives 92] before the transmission. Therefore, the password cannot be restored if the user forgets the password. The only way to recover the access to the system is to ask the system administrator to install the MD5-hash of a new password. Later on, the user can identify himself to the system with his username and password[2].

- **Shared Access:** In clinical practice a single patient is looked after by many persons. Hence, users of PEAKS have to be able to share the access to their patients. Users are able to work together in a group of users. The data of each individual user can be shared with other users.

  New recordings for the patients of another user can be created, and the properties of the patient, like weight, height, and disorder specific characteristics, can be altered. Deletion of patients who were created by other users is not permitted. Patients may only be deleted by the user who actually created the patient entry. The recordings of the patient will not be visible any longer to other users.

  To gain or remove privileges, the users have to contact the administrator. Users are not allowed to share their data themselves due to the layer 8 problem [Zimme 80].

[2] For security reasons both should be memorized by the users but not written down, because an attacker might gain access to the written copy of the login data.

4.2 Mathematical and Algorithmic Concepts

This section describes the necessary mathematical and algorithmic concepts which are employed in the PEAKS system. First, the methods to measure agreement are discussed since some of the measures form the basis of the later algorithms. Three different methods for the computation of the correlation are presented: Pearson's, Spearman's, and Cohen's method. Furthermore, two other parameters which are commonly used in the literature — Kappa and Alpha — are presented and their advantages and disadvantages are discussed.

Next, an introduction to Support Vector Machines is given in Chapter 4.2.2. The section explains Support Vector Classification and Support Vector Regression.

For the feature extraction and the visualization, dimensionality reduction techniques are required. Therefore, the Principal Component Analysis, the Linear Discriminant Analysis, and the Sammon Mapping are presented.

Speech processing is a fundamental part in the analysis of pathologic speech. The relevant methods which are discussed are feature extraction, speaker recognition, acoustic modeling, language modeling, decoding, prosodic analysis and the pronunciation assessment.

At the end of this section, some techniques to cope with the effects of different ages are described: Maximum Likelihood Linear Regression and Vocal Tract Length Normalization.

4.2.1 Measurements of Agreement

Since there are many scales and evaluation methods (cf. Chapter 3.1), there are also many methods to measure the agreement between different raters. A very robust and simple method to compute the agreement between different raters are correlation methods. These are especially useful if a quantitative evaluation method was employed since correlation measures are also defined on e.g. direct magnitude estimation scales.

If a nominal scale, e.g. a Likert-scale (cf. Table 3.1), was employed, Kappa and Alpha values can often be found in the literature to demonstrate the agreement. In the following these measures are discussed.

Correlation and Regression

Correlation coefficients and regression were developed in the biological sciences in order to proof evolutionary theories. Pearson was a student of Galton—a cousin of Charles Darwin. He developed the product-moment correlation coefficient in order to show the connection between parent and children generations in size and weight [Pears 96]. Today, correlation and regression have become very powerful tools in order to show whether a set of variables is connected or not. Note that correlation can only grasp the connection between two random variables but not clarify which variable is the cause and which one is the effect.

- **Correlation Coefficients:** The original correlation and regression as presented in [Pears 01] are computed using two variables $\mathcal{X}$ and $\mathcal{Y}$, where x_i and

y_i are the respective values which correspond to the same observation. The correlation $r_{\mathcal{X}\mathcal{Y}}$ between both variables is now computed as

$$\begin{aligned} r_{\mathcal{X}\mathcal{Y}} &= \frac{\sigma^2_{\mathcal{X}\mathcal{Y}}}{\sigma_{\mathcal{X}}\sigma_{\mathcal{Y}}} \qquad \text{with} && (4.1)\\ \sigma_{\mathcal{X}\mathcal{Y}} &= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_{\mathcal{X}})(y_i - \mu_{\mathcal{Y}})} && (4.2)\\ \sigma_{\mathcal{X}} &= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_{\mathcal{X}})^2} && (4.3)\\ \mu_{\mathcal{X}} &= \frac{1}{N}\sum_{i=1}^{N} x_i, && (4.4) \end{aligned}$$

where $\sigma_{\mathcal{X}}$ is the standard deviation of $\mathcal{X}$, $\sigma_{\mathcal{X}\mathcal{Y}}$ the covariance between $\mathcal{X}$ and $\mathcal{Y}$, $\mu_{\mathcal{X}}$ the mean of $\mathcal{X}$, and N the total number of observed pairs of both variables. $\sigma_{\mathcal{Y}}$ and $\mu_{\mathcal{Y}}$ are defined analog to the respective values of $\mathcal{Y}$.

The correlation can be interpreted as a normalized covariance between both variables. If the variables are not correlated, i.e. $r_{\mathcal{X}\mathcal{Y}} = 0$, both variables are independent of each other, i.e. $\sigma_{\mathcal{X}\mathcal{Y}} = 0$. The extreme values of $r_{\mathcal{X}\mathcal{Y}}$ are 1 and -1. If $r_{\mathcal{X}\mathcal{Y}}$ is close to 1, both variables are directly proportional, i.e. high values of $\mathcal{X}$ correspond to high values of $\mathcal{Y}$ and vice versa. If $r_{\mathcal{X}\mathcal{Y}}$ is near -1 both variables are inversely proportional i.e. high values of $\mathcal{X}$ correspond to low values of $\mathcal{Y}$ and vice versa. In the literature $r_{\mathcal{X}\mathcal{Y}}$ is often called r for simplicity reasons if the variables which are concerned can be inferred from the context. This procedure is used for this work as well.

Regression addresses the problem of the prediction of a target value y_i from an observed value x_i which is also called prediction value. The linear regression solves this by the computation of y_i using a linear transformation of x_i:

$$y_i = c_1 x_i + c_0 + \epsilon_i, \tag{4.5}$$

where c_1 is the slope of the regression, c_0 the intercept, and ϵ_i the error which cannot be explained linearly. The slope c_1 and intercept c_0 which have the least square error ϵ_{LSE} [Pears 01] are calculated as

$$\begin{aligned} c_1 &= \frac{\sigma_{\mathcal{X}\mathcal{Y}}}{\sigma_{\mathcal{X}}} && (4.6)\\ c_0 &= \mu_{\mathcal{Y}} - c_1\mu_{\mathcal{X}} \qquad \text{with} && (4.7)\\ \epsilon_{\text{LSE}} &= \sum_{i=1}^{N}\epsilon_i^2. && (4.8) \end{aligned}$$

The predicted values $\hat{y}_i$ which have the minimum square error are obtained by

$$\hat{y}_i = c_1 x_1 + c_0. \tag{4.9}$$

Note that the correlation between $\mathcal{X}$ and $\mathcal{Y}$ can also be formulated as the correlation between $\hat{\mathcal{Y}}$ and $\mathcal{Y}$ since $\hat{\mathcal{Y}}$ was created by linear transformation of $\mathcal{X}$:

$$r_{\hat{\mathcal{Y}}\mathcal{Y}} = \frac{\sigma^2_{\hat{\mathcal{Y}}\mathcal{Y}}}{\sigma_{\hat{\mathcal{Y}}}\sigma_{\mathcal{Y}}} = \frac{c_1\sigma^2_{\mathcal{X}\mathcal{Y}}}{c_1\sigma_{\mathcal{X}}\sigma_{\mathcal{Y}}} = \frac{\sigma^2_{\mathcal{X}\mathcal{Y}}}{\sigma_{\mathcal{X}}\sigma_{\mathcal{Y}}} = r_{\hat{\mathcal{X}}\mathcal{Y}}, \tag{4.10}$$

i.e. the normalized covariance $\frac{\sigma^2_{\hat{\mathcal{Y}}\mathcal{Y}}}{\sigma_{\hat{\mathcal{Y}}}}$ is the same as $\frac{\sigma^2_{\mathcal{X}\mathcal{Y}}}{\sigma_{\mathcal{X}}}$.

According to Spearman the correlation of the actual values after Pearson is very sensitive to outliers [Spear 04]. Therefore, he proposed to project the actual values onto their rank within the respective variable:

$$x_{i,\text{rank}} = f_{\text{rank}}(x_i) \tag{4.11}$$

where $f_{\text{rank}}(x_i)$ determines the rank of x_i. Spearman's rank correlation $\rho_{\mathcal{X}\mathcal{Y}}$ is computed as

$$\rho_{\mathcal{X}\mathcal{Y}} = r_{\mathcal{X}_{\text{rank}}\mathcal{Y}_{\text{rank}}}. \tag{4.12}$$

Especially for the evaluation of perceptive ratings, Spearman's correlation was shown to be more reliable. Furthermore, data with large outliers might increase Pearson's correlation artificially. Spearman's rank correlation is more robust to these outliers. In normally distributed data, r and ρ lie in the same range.

- **Multiple Regression / Correlation Analysis:** According to [Cohen 83] the problem formulated in Equation 4.5 can also be formulated for a prediction of y_i using a multidimensional vector $\boldsymbol{x}_i$ with n dimensions:

$$y_i = c_n x_{n,i} + c_{n-1}x_{n-1,i} + \ldots + c_1 x_{1,i} + c_0 + \epsilon_i \tag{4.13}$$

This can be rearranged to matrix annotation with the vectors $\boldsymbol{y}$ containing all target values and $\boldsymbol{c}$ with all prediction parameters to

$$\boldsymbol{y} = \boldsymbol{c}^\top \boldsymbol{X} \tag{4.14}$$

where $\boldsymbol{X}$ is the data matrix containing the vectors $\boldsymbol{x}_i$ as column vectors plus an additional entry 1 for the intercept of the regression. The prediction parameter vector $\boldsymbol{c}$ can now be computed as

$$\boldsymbol{c}^\top = \boldsymbol{y}\boldsymbol{X}^* \tag{4.15}$$

where $\boldsymbol{X}^*$ is the Moore-Penrose pseudo-inverse of $\boldsymbol{X}$ [Moore 20, Penro 55] which computes the best approximation of the inverse according to the least square error using singular value decomposition. Thus, the predictions of y_i can now be computed as

$$\hat{y}_i = \boldsymbol{c}^\top \begin{pmatrix} \boldsymbol{x}_i \\ 1 \end{pmatrix}. \tag{4.16}$$

In order to compute the correlation between the multidimensional input variable $\mathcal{X}$ and $\mathcal{Y}$, the one-dimensional version of the correlation can be employed since Equation 4.10 is still valid. Thus, the multi-correlation is

$$R_{\mathcal{X}\mathcal{Y}} = r_{\hat{\mathcal{Y}}\mathcal{Y}}. \tag{4.17}$$

- **Significance Tests:** In order to determine the significance of a result at a certain level of chance α, the probability that the result which was observed is pure coincidence p_α is computed. This probability can be formulated as the integral of the probability density function of r from $1-\alpha$ to 1. The probability density function of r, however, is dependent on the number of test cases N, the degrees of freedom $f_{\text{df}} = N - 2$, and on the actual value of r.

  Instead of the computation of all possible density functions of r, the density function is usually transformed to a density function with well-known statistics. The density function of r can be mapped to Student's distribution. According to [Stang 71] the integral in Student's distribution t is computed as:

  $$t = \frac{r}{\sqrt{1-r^2}}\sqrt{f_{\text{df}}} \tag{4.18}$$

  Now, the significance decision can be done according to the following rules:

  $$|r| \begin{cases} = 0 & \text{if} \quad t < t_{f_{\text{df}};1-\alpha} \\ \\ > 0 & \text{if} \quad t > t_{f_{\text{df}};1-\alpha} \end{cases} \tag{4.19}$$

  The statistics for $t_{f_{\text{df}};1-\alpha}$ can be reviewed in most text books on statistics [Stang 71].

  In order to compute whether two correlations r_1 and r_2 are significantly different, the distributions of r_1 and r_2 have to be transformed to a single distribution of known statistics. In this case, both can be mapped to a standard normal distribution [Stang 71]. The integral u between $1-\alpha$ and 1 is now computed as

  $$u = \frac{z(r_1) - z(r_2)}{\sqrt{\frac{1}{N_1-3} + \frac{1}{N_2-3}}} \quad \text{with} \tag{4.20}$$

  $$z(r) = \frac{1}{2} ln \left(\frac{1+r}{1-r}\right) \tag{4.21}$$

  where N_1 is the number of cases in r_1, and N_2 is the number of cases in r_2. The decision is now performed according to:

  $$r_1 \begin{cases} = r_2 & \text{if} \quad |u| < u_{1-(\alpha/2)} \\ \\ \neq r_2 & \text{if} \quad |u| > u_{1-(\alpha/2)} \end{cases} \tag{4.22}$$

  Again the values of $u_{1-(\alpha/2)}$ can be reviewed in common statistics text books [Stang 71]. For small numbers of test subjects and high correlations, the approximations for the significance might be incorrect. In [Ogus 07], significance tables for $r > 0.2$ and sample sizes of 18 or less are given.

Kappa

Cohen's Kappa (here denoted as $\mathcal{K}$) reports the agreement between two raters A and B [Cohen 60]. In order to compensate that two raters might be in concordance by chance, the actually observed agreement $\pi^{(A,B)}$ is normalized by the chance agreement $\pi_\epsilon^{(A,B)}$ between both raters. $\mathcal{K}$ is only applicable to nominal data, i.e., data which is separated into a fixed number of classes or categories.

$$\mathcal{K}^{(\mathcal{A},\mathcal{B})} = \frac{\pi^{(\mathcal{A},\mathcal{B})} - \pi_\epsilon^{(\mathcal{A},\mathcal{B})}}{1 - \pi_\epsilon^{(\mathcal{A},\mathcal{B})}} \tag{4.23}$$

According to Krummenauer [Krumm 99], $\pi^{A,B}$ can be computed as

$$\pi^{(A,B)} = \sum_{\kappa=1}^{K}\sum_{\iota=1}^{K} \pi_{\kappa,\iota}^{(A,B)} \tag{4.24}$$

where $\pi_{\kappa,\iota}^{(A,B)}$ denotes the probability that a subject is assigned to class Ω_κ by rater A and to category Ω_ι by rater B, and K is the total number of classes.

The chance agreement $\pi_\epsilon^{(A,B)}$ is determined as

$$\pi_\epsilon^{(A,B)} = \sum_{\kappa=1}^{K}\sum_{\iota=1}^{K} \pi_\kappa^{(A)} \cdot \pi_\iota^{(B)} \tag{4.25}$$

where $\pi_\kappa^{(A)}$ and $\pi_\iota^{(B)}$ are the probabilities for raters A and B to choose the classes Ω_κ and Ω_ι at all.

Since the probabilities $\pi_{\kappa,\iota}^{(A,B)}$ and $\pi_\kappa^{(A)}$ are usually not available, they have to be estimated from the data set:

$$\hat{\pi}_{\kappa,\iota}^{(A,B)} = \frac{\#(\kappa,\iota)}{N} \quad \text{and} \tag{4.26}$$

$$\hat{\pi}_\kappa^{(A)} = \frac{\#(\kappa)}{N} \tag{4.27}$$

$\#(\kappa,\iota)$ denotes the number of confusions between Ω_κ and Ω_ι and $\#(\kappa)$ the total number of occurrences of Ω_κ.

An extension to Cohen's Kappa is to weigh the disagreement between two raters using a weighting function [Fleis 69]. In [Cicch 76] an absolute and a squared weighting function $w_{\kappa,\iota}^{(A,B)} \in [0,1]$ are proposed:

$$w_{\kappa,\iota,\text{abs}}^{(A,B)} = 1 - \left|\frac{\kappa-\iota}{K-1}\right| \tag{4.28}$$

$$w_{\kappa,\iota,\text{sqr}}^{(A,B)} = 1 - \left(\frac{\kappa-\iota}{K-1}\right)^2 \tag{4.29}$$

Either of both can be applied in the following extension of $\mathcal{K}(\mathcal{A},\mathcal{B})$:

$$\pi^{(A,B)} = \sum_{\kappa=1}^{K}\sum_{\iota=1}^{K} \pi_{\kappa,\iota}^{(A,B)} \cdot w_{\kappa,\iota}^{(A,B)} \tag{4.30}$$

$$\pi_\epsilon^{(A,B)} = \sum_{\kappa=1}^{K}\sum_{\iota=1}^{K} \pi_\kappa^{(A)} \cdot \pi_\iota^{(B)} \cdot w_{\kappa,\iota}^{(A,B)} \tag{4.31}$$

Note that Equation 4.23 still applies for the computation of $\mathcal{K}^{(\mathcal{A},\mathcal{B})}$.

A multi-rater version of Kappa $\mathcal{K}_{\mathrm{DF}}$ is presented in [Fleis 71, Davie 82]. Therefore, the Kappa between two raters is normalized by the chance agreement between both of them and then combined for all raters R according to:

$$\mathcal{K}_{\mathrm{DF}} = \frac{\sum_{\mathcal{A}=\infty}^{\mathcal{R}} \sum_{\mathcal{B}\neq\mathcal{A}}^{\mathcal{R}} [1 - \pi_{\epsilon}^{(\mathcal{A},\mathcal{B})}] \cdot \mathcal{K}^{(\mathcal{A},\mathcal{B})}}{\sum_{\mathcal{A}=\infty}^{\mathcal{R}} \sum_{\mathcal{B}\neq\mathcal{A}}^{\mathcal{R}} [1 - \pi_{\epsilon}^{(\mathcal{A},\mathcal{B})}]} \tag{4.32}$$

Note that this combination of the binary Kappas $\mathcal{K}^{(\mathcal{A},\mathcal{B})}$ is possible for the weighted and for the non-weighted case.

The major advantage of the weighted multi-rater Kappa is that the chance factor of the agreement is leveled out. However, it is only applicable for nominal classes and it can not handle missing data. Therefore, it is only applied in this work if both constraints are met.

Alpha

In order to handle missing data, Krippendorff's Alpha (here denoted as $\mathcal{A}$) models the disagreement D_0 and the expected disagreement D_ϵ between the raters instead of the agreement [Kripp 03]:

$$\mathcal{A} = 1 - \frac{D_0}{D_\epsilon} \quad \text{with} \tag{4.33}$$

$$D_0 = \frac{1}{N} \sum_{\kappa}^{K} \sum_{\iota\neq\kappa}^{K} \upsilon_{\kappa\iota} \cdot \delta_{\kappa\iota}^2 \tag{4.34}$$

$$D_\epsilon = \frac{1}{N(N-1)} \sum_{\kappa}^{K} \sum_{\iota\neq\kappa}^{K} \#(\kappa) \cdot \#(\iota) \cdot \delta_{\kappa\iota}^2 \tag{4.35}$$

$$\delta_{\kappa\iota}^2 = (\kappa - \iota)^2 \tag{4.36}$$

where $\upsilon_{\kappa\iota}$ is the observable disagreement for Ω_κ and Ω_ι, $\mathcal{A}$ is 0 if the expected disagreement D_ϵ equals the observed disagreement D_0.

For the computation of $\upsilon_{\kappa\iota}$ the number of observed confusions $\#_s(\kappa,\iota)$ between class Ω_κ and class Ω_ι for each subject s has to be determined. Then $\upsilon_{\kappa\iota}$ is computed as

$$\upsilon_{\kappa\iota} = \sum_{s} \frac{\#_s(\kappa,\iota)}{\#(s) - 1} \tag{4.37}$$

where $\#(s)$ is the number of ratings which were given to subject s.

While Alpha solves the problem of missing data by the computation of the disagreement, it is still just defined on nominal classes. In principle, real valued classes may be used. However, the values of all classes must be in the same range in order to get a feasible Alpha.

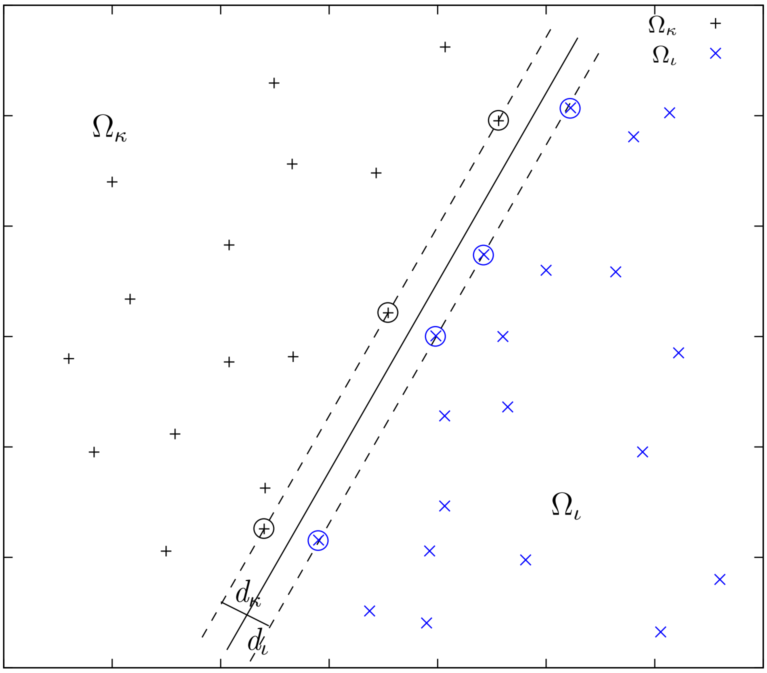


Figure 4.4: In order to separate two classes from each other, an optimal plane can be found. In the example the Support Vectors are marked with circles.

4.2.2 Support Vector Machines

The theory of Support Vectors (SVs) goes back to Vapnik who developed the *Generalized Portrait* algorithm in the sixties of the last century in Russia [Vapni 63]. The Support Vector Machines (SVMs) used today are a further generalization of this algorithm. One of the strong points of SVMs is that they still have a good generalization even if the training set is quite small and that they model outliers well. In the following, SV classification and SV regression are described.

Support Vector Classification

Support Vector Classification [Schol 97] tries to find a plane which separates two classes Ω_κ and Ω_ι from each other. Therefore, we don't need to remember all N observations in the training set. Only a small number of observations is really important for the classification task. As shown in Figure 4.4, only those vectors which are marked by a circle are important for the decision. We want to determine $(\boldsymbol{w}, b)$ which define the plane that separates the feature space of the individual feature vectors $\boldsymbol{x}_i$. In mathematical terms this problem can be formulated as

$$(\boldsymbol{w}^\top \boldsymbol{x}_i) + b) \geq 1, \qquad \text{for } y = 1 \tag{4.38}$$

$$(\boldsymbol{w}^\top \boldsymbol{x}_i) + b) \leq -1, \qquad \text{for } y = -1 \tag{4.39}$$

where $y_i = 1$ denotes that $\boldsymbol{x}_i$ is a member of class Ω_κ and $y_i = -1$ assigns $\boldsymbol{x}_i$ to class Ω_ι. A set of inequalities can be formed:

$$y_i((\boldsymbol{w}^\top \boldsymbol{x}_i) + b) \geq 1, \quad i = 1, \ldots, N \tag{4.40}$$

If we just consider the points for which the equality in equations Eq. 4.38 and Eq. 4.39 hold, we get points which are located on two hyperplanes $H_\kappa \; : \; \boldsymbol{w}^\top \boldsymbol{x}_i + b = 1$ and $H_\iota \; : \; \boldsymbol{w}^\top \boldsymbol{x}_i + b = -1$. The plane H_κ has a perpendicular distance to the origin of $|1 - b|/||w||$ where $||w||$ is the norm of vector $\boldsymbol{w}$ and H_ι has a distance of $|-1-b|/||\boldsymbol{w}||$ to the origin. Thus, the distances between the optimal hyperplane and H_κ and H_ι are $d_\kappa = d_\iota = 1/||\boldsymbol{w}||$. The margin between both planes is simply $2/||\boldsymbol{w}||$. Hence minimizing $||\boldsymbol{w}||$ subject to the constraints from Eq. 4.40 gives the normal vector of the hyperplane which separates the feature space optimally.

The optimization problem can now be formulated as a Lagrangian using the Lagrange multipliers α_i. The primary Lagrangian L_P is formed as

$$L_P = \frac{1}{2}||\boldsymbol{w}||^2 - \sum_i \alpha_i y_i (\boldsymbol{w}^\top \boldsymbol{x}_i + b) + \sum_i \alpha_i \tag{4.41}$$

For optimality the gradient of L_P needs to vanish for $\boldsymbol{w}$ and b. Setting the respective derivatives to zero yields the following equations:

$$\boldsymbol{w} = \sum_i \alpha_i y_i \boldsymbol{x}_i \tag{4.42}$$

$$0 = \sum_i \alpha_i y_i \tag{4.43}$$

Note that Eq. 4.42 states that the normal vector of the optimal hyperplane is composed of a weighted sum of all feature vectors which a non-zero α_i, i.e. the Support Vectors. Equations Eq. 4.42 and Eq.4.43 can now be plugged into Eq. 4.41 which yields the so-called Wolfe dual [Fletc 87]:

$$L_D = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^\top \boldsymbol{x}_j \tag{4.44}$$

In order to allow misclassifications (cf. Figure 4.5), a slack variable $\xi_i \geq 0$ is introduced. Equation 4.40 becomes

$$y_i((\boldsymbol{w} \cdot \boldsymbol{x}_i) + b) \geq 1 - \xi_i, \quad i = 1, \ldots, N. \tag{4.45}$$

According to [Burge 98] this yields the following extended primary Lagrangian with Lagrange multipliers α_i and η_i:

$$L_P = \frac{1}{2}||w||^2 + C \sum_i \xi_i - \sum_i \alpha_i [y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) - 1 + \xi_i] - \sum_i \eta_i \xi_i \tag{4.46}$$

C is a penalty parameter which is to be chosen by the user. The larger C is the higher is the penalty to errors. Using Karush-Kuhn-Tucker conditions [Fletc 87], Eq. 4.46 can be reduced again to Eq. 4.44. Another great advantage is that SVMs can be easily extended with a so-called kernel function which allows to use non-linear separation functions [Burge 98]. Without loss of generality, this procedure can be extended to more classes. Therefore, all planes between the two respective classes Ω_i and Ω_j have to be determined in the same manner.

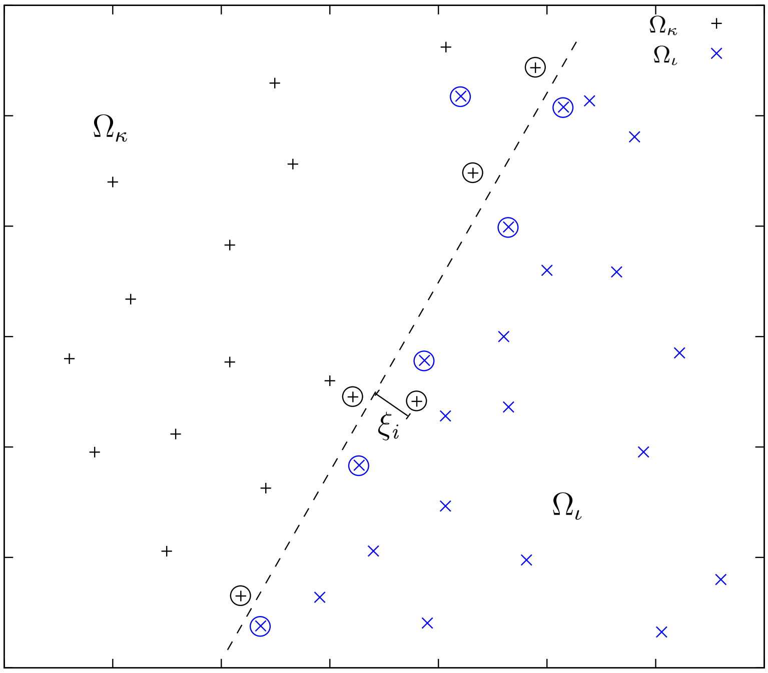


Figure 4.5: In the non-separable case, misclassifications have to be allowed. Therefore, slack variables ξ_i are introduced.

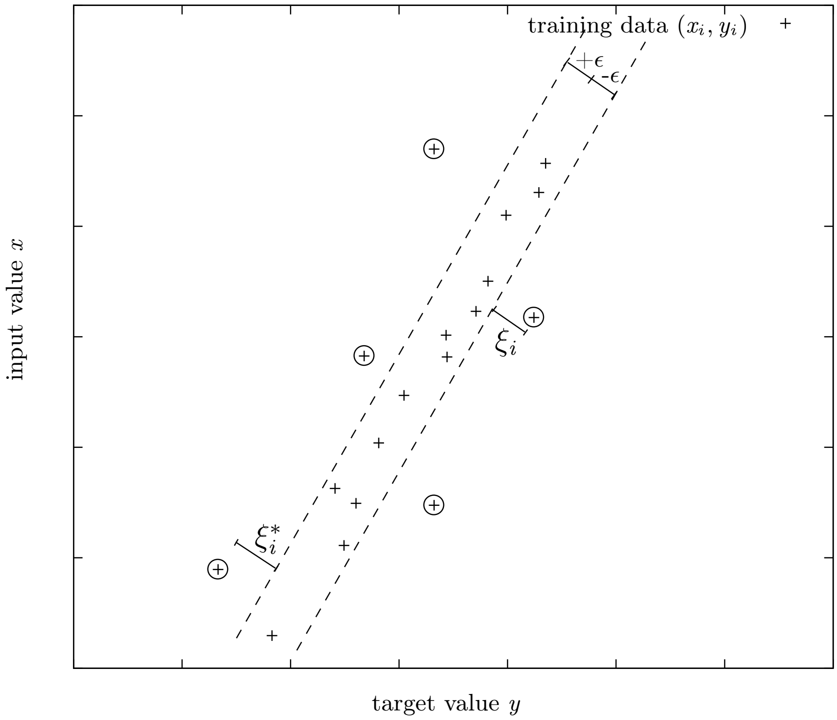


Figure 4.6: Support Vector Regression finds a function that has at most ϵ deviation from the targets y_i. In order to allow deviations larger than ϵ, a slack variable ξ_i is introduced again. Note that the support vectors are outside the ϵ-tube.

Support Vector Regression

In order to approximate an arbitrary function, *Support Vector Regression* (SVR) [Smola 98] can be applied. Its goal is to compute an estimate value $\hat{y}_i$ for each of the N feature vectors $\boldsymbol{x}_i$ which deviate at most ϵ from the original target value y_i. This leads to the following equation:

$$\hat{y}_i = \boldsymbol{w}^\top \boldsymbol{x}_i + b. \tag{4.47}$$

The variables $\boldsymbol{w}$ and b are found by solving the problems

$$y_i - (\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \leq \epsilon \quad \text{and} \quad (\boldsymbol{w} \cdot \boldsymbol{x}_i + b) - y_i \leq \epsilon. \tag{4.48}$$

To allow deviations greater than ϵ, slack variables ξ_i and ξ_i^* are introduced again. Equation 4.48 can be rewritten to

$$y_i - (\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \leq \epsilon + \xi_i \quad \text{and} \quad (\boldsymbol{w} \cdot \boldsymbol{x}_i + b) - y_i \leq \epsilon + \xi_i^*. \tag{4.49}$$

In order to constrain the type of the vector $\boldsymbol{w}$, we postulate *flatness*. One way to achieve this is to minimize its norm $||\boldsymbol{w}||$. We end in the following minimization problem:

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_i(\xi_i + \xi_i^*) \\ \text{subject to} & \begin{cases} y_i - (\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \leq \epsilon + \xi_i \\ (\boldsymbol{w} \cdot \boldsymbol{x}_i + b) - y_i \leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \end{array} \tag{4.50}$$

Similar as for the case of SVMs, a primal Lagrangian can be formulated introducing Lagrange multipliers α_i, α_i^*, η_i, and η_i^* in order to solve this problem.

$$\begin{aligned} L_P = \frac{1}{2}||w||^2 + C\sum_i(\xi_i + \xi_i^*) - \sum_i \alpha_i(\epsilon + \xi_i - y_i + \boldsymbol{w}^\top \boldsymbol{x}_i + b) \\ - \sum_i \alpha_i^*(\epsilon + \xi_i^* + y_i - \boldsymbol{w}^\top \boldsymbol{x}_i - b) - \sum_i(\eta_i \xi_i + \eta_i^* \xi_i^*) \end{aligned} \tag{4.51}$$

Again, C denotes a penalty parameter to be chosen by the user. The saddle point condition of a minimum requires the derivative of L_P to vanish for the primal variables $\boldsymbol{w}$, b, ξ_i, and ξ_i^*. Therefore, partial derivation of L_P yields the following equations:

$$0 = \sum_i(\alpha_i^* - \alpha_i) \tag{4.52}$$

$$\boldsymbol{w} = \sum_i(\alpha_i - \alpha_i^*)\boldsymbol{x}_i \tag{4.53}$$

$$0 = C - \alpha_i^{(*)} - \eta_i^{(*)} \tag{4.54}$$

By substitution of the equations Eq. 4.52 to Eq. 4.54 in Eq. 4.51, the following optimization problem is obtained:

$$\begin{aligned} &\text{maximize} && \begin{cases} -\dfrac{1}{2}\displaystyle\sum_{i,j}(\alpha_i-\alpha_j)(\alpha_i^*-\alpha_j^*)\boldsymbol{x}_i^\top\boldsymbol{x}_j \\ -\epsilon\displaystyle\sum_i(\alpha_i-\alpha_i^*)+\sum_i y_i(\alpha_i-\alpha_i^*) \end{cases} \\ &\text{subject to} && \begin{cases} \displaystyle\sum_i(\alpha_i-\alpha_i^*)=0 \\ \alpha_i,\alpha_i^*\in[0,C] \end{cases} \end{aligned} \tag{4.55}$$

Note that the Lagrange multipliers η_i and η_i^* are eliminated in the derivation of Eq. 4.55. According to [Smola 98] the constraint $\alpha_i\alpha_i^*=0$ has to be met. Thus, there can never be a set of variables α_i and α_i^* which both are nonzero at the same time. Furthermore, α_i and α_i^* are zero if $|\hat{y}_i-y_i|\le\epsilon$. Therefore, Support Vectors can only be found outside the ϵ-tube (cf. Figure 4.6). With Eq. 4.53 the prediction of $\hat{y}_i$ from Eq. 4.47 can now be written without the actual weight vector $\boldsymbol{w}$:

$$\hat{y}_i=\left[\sum_j(\alpha_j-\alpha_j^*)\boldsymbol{x}_j\right]^\top\boldsymbol{x}_i+b \tag{4.56}$$

4.2.3 Dimension Reduction Techniques

In pattern recognition the reduction of the dimension of a feature space is desirable. This is useful to extract the important information in terms of classification, regression, or visualization. Basically this can be done by linear and non-linear methods.

All linear methods can be written as a matrix product [Niema 03, p.154]

$$\hat{\boldsymbol{x}}=\boldsymbol{\Phi}\boldsymbol{x} \tag{4.57}$$

where $\hat{\boldsymbol{x}}$ denotes the transformed feature vector of dimension n', $\boldsymbol{\Phi}$ the transformation matrix, and $\boldsymbol{x}$ the original feature vector of dimension n. The linear dimension reduction techniques presented in the following are the Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Feature Selection (FS). They differ just in the configuration of the matrix $\boldsymbol{\Phi}$.

As non-linear dimensionality reduction techniques the Sammon Mapping, which tries to preserve the topology of the high-dimensional space, and an extension to the Sammon Mapping, which incorporates additional prior knowledge on the mapping, are presented.

Principal Component Analysis

The PCA finds the principal components which contain the most variance and project the data accordingly (cf. Figure 4.7). The optimization criterion for the PCA is the mean square distance e_{MSD} between all vectors. With N being the number of vectors, the e_{MSD} between all vectors is defined as

$$\epsilon_{MSD}=\frac{1}{N^2}\sum_{i=1}^N\sum_{j=1}^N\left(\hat{\boldsymbol{x}}_i-\hat{\boldsymbol{x}}_j\right)^\top\left(\hat{\boldsymbol{x}}_i-\hat{\boldsymbol{x}}_j\right). \tag{4.58}$$

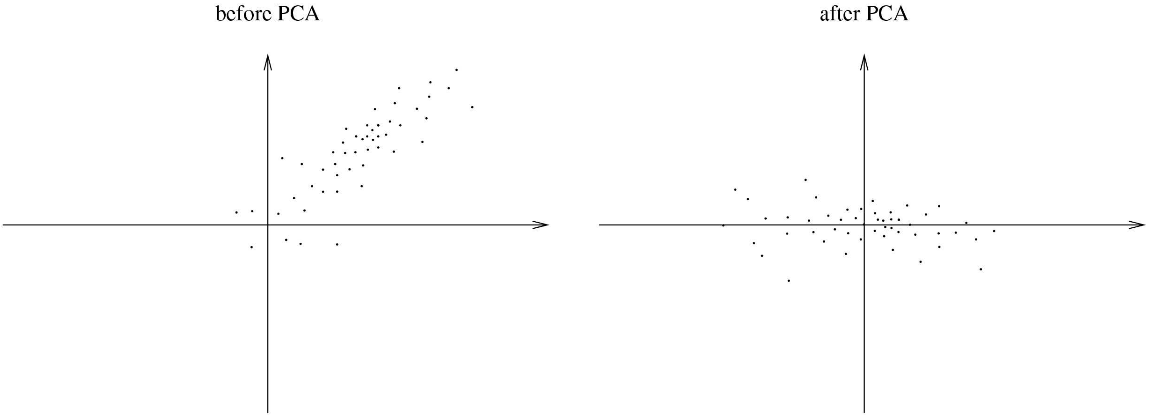


Figure 4.7: The PCA projects the features onto their principal components. The axis which contains the most variance is projected onto the first axis.

The problem of computing optimized features can be reduced to the maximization of a quadratic form [Niema 03, p. 206]. $\boldsymbol{\Phi}_{\text{PCA}}$ is obtained by computing the eigenvectors $\boldsymbol{\varphi}_\nu$ from a suitable core matrix $\boldsymbol{Q}$:

$$\boldsymbol{Q}\ \boldsymbol{\varphi}_\nu = \lambda_\nu\ \boldsymbol{\varphi}_\nu \tag{4.59}$$

With λ_ν being the eigenvalues of $\boldsymbol{Q}$, the n largest eigenvectors $\boldsymbol{\varphi}_\nu$ can be computed. Thus, the transformation matrix $\boldsymbol{\Phi}_{\text{PCA}}$ is of the from

$$\boldsymbol{\Phi}_{\text{PCA}} = \begin{pmatrix} \boldsymbol{\varphi}_1^\top \\ \boldsymbol{\varphi}_2^\top \\ \vdots \\ \boldsymbol{\varphi}_n^\top \end{pmatrix} \tag{4.60}$$

Using the covariance matrix $\boldsymbol{\Sigma}$ and the mean vector $\boldsymbol{\mu}$ of all data points, a suitable core matrix $\boldsymbol{Q}$ is determined [Maier 08a]:

$$\begin{aligned} \boldsymbol{Q} &= \boldsymbol{\Sigma} - \boldsymbol{\mu}\boldsymbol{\mu}^\top \\ \boldsymbol{\Sigma} &= \frac{1}{N}\sum_{j=1}^{N} \boldsymbol{x}_j \boldsymbol{x}_j^\top, \qquad \boldsymbol{\mu} = \frac{1}{N}\sum_{j=1}^{N} \boldsymbol{x}_j \end{aligned} \tag{4.61}$$

The proof to show that this core matrix $\boldsymbol{Q}$ corresponds to ϵ_{MSD} is based on substituting Eq. 4.58 into Eq. 4.57 [Maier 05b]. The resulting transformation matrix projects the axis which contains the most variance of the data onto the first component and the axis containing the second most variance onto the second component and so on (cf. Eq. 4.60).

Linear Discriminant Analysis

For the LDA another criterion is chosen which takes the class membership into account. Each vector $\boldsymbol{x}_\kappa$ is member of a class Ω_κ. K denotes the number of classes while

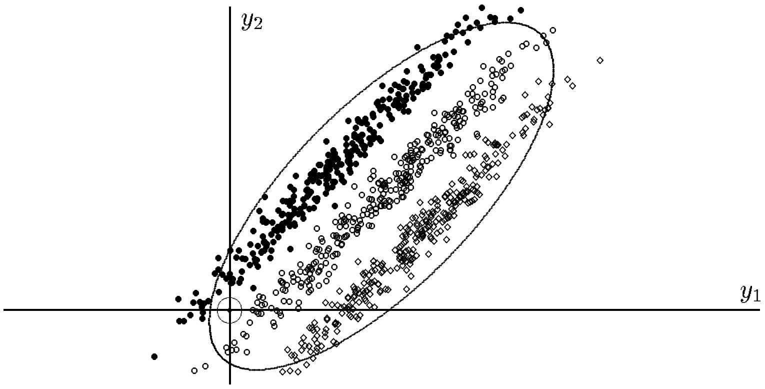


Figure 4.8: A distribution of three classes which cannot be transformed by the PCA optimally if the dimension is reduced – the "adidas problem" [Schuk 95, p. 116].

N_κ reflects the number of elements in class Ω_κ in the training set. Thus, $N = \sum_{\kappa=1}^{k} N_\kappa$ is the total number of feature vectors.

As described in [Stemm 05, p.41], from this information two covariance matrices can be computed.

$$\begin{aligned}
\boldsymbol{W}_{\mathrm{LDA}} &= \frac{1}{N}\sum_{\kappa=1}^{K}\sum_{j=1}^{N_\kappa} \boldsymbol{x}_{\kappa_j}\boldsymbol{x}_{\kappa_j}^\top \qquad (4.62)\\
\boldsymbol{B}_{\mathrm{LDA}} &= \frac{1}{N}\sum_{\kappa=1}^{K} N_\kappa \boldsymbol{\mu}_\kappa \boldsymbol{\mu}_\kappa^\top - \left(\frac{1}{N}\sum_{\kappa=1}^{K} N_\kappa \boldsymbol{\mu}_\kappa\right)\left(\frac{1}{N}\sum_{\kappa=1}^{K} N_\kappa \boldsymbol{\mu}_\kappa\right)^\top, \quad \text{where}\\
\boldsymbol{\mu}_\kappa &= \frac{1}{N_\kappa}\sum_{j=1}^{N_\kappa} \boldsymbol{x}_{\kappa_j}
\end{aligned}$$

$\boldsymbol{W}_{\mathrm{LDA}}$ states the scatter within one class while $\boldsymbol{B}_{\mathrm{LDA}}$ states the scatter between the classes. The complete covariance matrix $\boldsymbol{\Sigma}$ can be obtained by the sum of both matrices.

An optimal transformation in terms of the LDA projects vectors of the same class very close together and vectors of different classes as far away as possible. We wish to maximize $\boldsymbol{B}_{\mathrm{LDA}}$ while minimizing $\boldsymbol{W}_{\mathrm{LDA}}$. Accordingly, we can formulate the following criterion:

$$sc(\boldsymbol{\Phi}_{\mathrm{LDA}}) = \frac{|\boldsymbol{\Phi}_{\mathrm{LDA}}\boldsymbol{B}_{\mathrm{LDA}}\boldsymbol{\Phi}_{\mathrm{LDA}}^\top|}{|\boldsymbol{\Phi}_{\mathrm{LDA}}\boldsymbol{W}_{\mathrm{LDA}}\boldsymbol{\Phi}_{\mathrm{LDA}}^\top|} \qquad (4.63)$$

In [Fukun 90] it is shown that this criterion can be maximized by setting the rows of $\boldsymbol{\Phi}_{\mathrm{LDA}}$ to the largest eigenvectors $\boldsymbol{\varphi}_\nu$ of the generalized eigenvector problem:

$$\boldsymbol{B}_{\mathrm{LDA}}\boldsymbol{\varphi}_\nu = \lambda_\nu \boldsymbol{W}_{\mathrm{LDA}}\boldsymbol{\varphi}_\nu \qquad (4.64)$$

The eigenvectors $\boldsymbol{\varphi}_\nu$ which are defined by the equation above are not orthonormal. The eigenvalues, however, λ_ν are always real and positive. Figure 4.8 shows a data set with different classes which are separated best by the second principal component. Transformation with PCA would project the axis with the highest variance onto the first component. If the dimension would be reduced, the class-separating information would be lost. The LDA transforms the data shown in Figure 4.8 onto its second principal component, which is the most discriminating one.

Feature Selection

The last linear dimension reduction technique presented here is the feature selection. The linear transformation $\boldsymbol{\Phi}_{\mathrm{FS},\nu}$ for the removal of a single feature x_ν is only composed of ones and zeros. $\boldsymbol{\Phi}_{\mathrm{FS},\nu}$ is a skewed $n \times (n-1)$ unity matrix. The column which corresponds to x_ν contains just zeros:

$$\boldsymbol{\Phi}_{\mathrm{FS},\nu}\boldsymbol{x} = \begin{pmatrix} 1 & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots\cdots & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_\nu \\ \vdots \\ x_q \\ \vdots \\ x_{n-1} \\ x_n \end{pmatrix} = \hat{\boldsymbol{x}} \tag{4.65}$$

x_q is an arbitrary entry in the feature vector different from x_ν. Removal of multiple features can be performed by the combination of several feature removal matrices $\boldsymbol{\Phi}_{\mathrm{FS},\nu}$ after each other.

While the actual removal is rather simple, the algorithms which actually select the features to be removed are manifold. All of these algorithms require the computation of a certain “quality” $G^{\mathcal{S}}$ of a subset of features $\mathcal{S}$ containing $N_{\mathcal{S}}$ elements. This $G^{\mathcal{S}}$ can be computed from different distance measures concerning the features in the subset, classification performance or estimates of the classification performance, confusion probabilities between the classes, regression analysis of the features to the class attribute, and many more [Hall 98, Niema 03, Clark 04, Cinca 04a].

In this work two quality criteria are inspected closely due to their direct application of regression: *Correlation-based Feature Subset* (CFS) and *Maximum R* (MAXR) selection. Although both of them seem quite similar, there are certain differences which result in different performance concerning the quality of the resulting regression.

The idea behind the CFS selection algorithm is to compute the correlation of a composite variable $\mathcal{X}^{\mathcal{S}}$ to an outside variable $\mathcal{Y}$ as the criterion for the quality. In [Ghise 64, p.182] a formulation of this correlation as a composition of the inter-correlations $r_{\mathcal{Y}x_i^{\mathcal{S}}}$ between the target variable $\mathcal{Y}$ and the $N_{\mathcal{S}}$ individual features $x_i^{\mathcal{S}}$ and the intra-correlations $r_{x_i^{\mathcal{S}}x_j^{\mathcal{S}}}$ is found:

$$r_{\mathcal{Y}\mathcal{X}^{\mathcal{S}}} = \frac{N_{\mathcal{S}}\overline{r_{\mathcal{Y}x_i^{\mathcal{S}}}}}{\sqrt{N_{\mathcal{S}} + N_{\mathcal{S}}(N_{\mathcal{S}}-1)\overline{r_{x_i^{\mathcal{S}}x_j^{\mathcal{S}}}}}} = G^{\mathcal{S}}_{\mathrm{CFS}} \tag{4.66}$$

$\overline{r}$ denotes the mean of the respective correlations. In [Hall 98], Eq. 4.66 is used to create a fast and efficient algorithm to select features which have a good correlation with the target variable. The computation is very efficient since the correlations between all variables just have to be computed once. After their computation the single correlations are stored in a lookup table which allows fast and easy access to the values.

However, the question arises why this computation is so simple while the computation of $R_{\mathcal{Y}\mathcal{X}^{\mathcal{S}}}$ from Eq. 4.17 would involve at least one matrix inversion. The answer is that the correlation determined with Eq. 4.66 is not weighted. Therefore, not the optimal weighting of the component variable is used in the CFS algorithm but the fusion of the unweighted components.

As one might suspect, the use of optimally weighted components in a least square error sense is able to preserve more information during the feature selection process. An algorithm which employs the multiple correlation coefficient $R_{\mathcal{Y}\mathcal{X}^{\mathcal{S}}}$ is described briefly in [Clark 04, p.34]. The MAXR algorithm simply sets

$$G^{\mathcal{S}}_{\text{MAXR}} = R_{\mathcal{Y}\mathcal{X}^{\mathcal{S}}} \tag{4.67}$$

to compute the quality of the feature subset $\mathcal{S}$.

The calculation of $R_{\mathcal{Y}\mathcal{X}^{\mathcal{S}}}$ in Eq. 4.67 is a crucial point for the application of the algorithm. The formulation of $R_{\mathcal{Y}\mathcal{X}^{\mathcal{S}}}$ in Eq. 4.17 involves a matrix inversion with a theoretical complexity of at least $\mathcal{O}(N_{\mathcal{S}}^3)$ (cf. [Press 92, Courr 05]) for each subset. This is quite expensive since implementations of the inverse matrix based on the QR decomposition, like the one in Weka [Witte 05], usually have a complexity of

$$\mathcal{O}^{\text{R-iter}} = \mathcal{O}(N_{\mathcal{S}}^2 N) \tag{4.68}$$

where N is the number of training vectors.

A faster approximation of R can be computed by gradient descent: Let $\boldsymbol{X}_{\mathcal{S}}$ be the data matrix which contains all features of subset $\mathcal{S}$. If $\mathcal{S}$ is of cardinality $n-1$, this $\boldsymbol{X}_{\mathcal{S}}$ can be computed by the multiplication of $\boldsymbol{\Phi}_{\text{FS},\nu}$ from Eq. 4.65 with $\boldsymbol{X}$ to remove feature number ν. The parameters $\boldsymbol{c}_{\mathcal{S}}$ can be computed according to Eqs. 4.14 and 4.15:

$$\boldsymbol{c}_{\mathcal{S}}^{\top} = \boldsymbol{y}\boldsymbol{X}_{\mathcal{S}}^{*} = \boldsymbol{y}\boldsymbol{X}^{*}\boldsymbol{\Phi}_{\text{FS},\nu}^{\top} = \boldsymbol{c}^{\top}\boldsymbol{\Phi}_{\text{FS},\nu}^{\top} \tag{4.69}$$

Note that $\boldsymbol{\Phi}_{\text{FS},\nu}^{\top}$ is the pseudo-inverse of $\boldsymbol{\Phi}_{\text{FS},\nu}$ since it is almost a diagonal matrix. This implies that the computationally very expensive matrix inversion has to be performed only once for all feature subsets $\mathcal{S}$. In order to refine the approximation further, a gradient descent can now be performed. The objective function of the descent is chosen as the sum of the square error of the prediction ϵ_{R}:

$$\epsilon_{\text{R}}(\boldsymbol{c}_{\mathcal{S}}) = \sum_{i=1}^{N} \left(\boldsymbol{c}^{\top}\boldsymbol{x}_i - y_i\right)^2 \tag{4.70}$$

Differentiation after each component c_j yields the following gradient function:

$$\frac{\delta\epsilon_{\text{R}}}{\delta c_j} = \sum_{i=1}^{N} \left(\boldsymbol{c}^{\top}\boldsymbol{x}_i - y_i\right) \cdot 2x_{i,j} \tag{4.71}$$

Using Eq. 4.69 as initialization for the gradient descent yields a quite good convergence behavior. In terms of complexity, this procedure surpasses all previous methods:

Since the sums of Eqs. 4.70 and 4.71 require just a single pass in each iteration, the complexity $\mathcal{O}^{\text{R-grad}}$ of this methods is

$$\mathcal{O}^{\text{R-grad}} = \mathcal{O}(N \cdot N_{\mathcal{S}} \cdot 2 \cdot C) \tag{4.72}$$

where C denotes a constant which corresponds to the number of iterations of the gradient descent. As soon as $\mathcal{O}^{\text{R-iter}} > \mathcal{O}^{\text{R-grad}}$, the feature selection should be performed with the gradient descent method in order to speed the feature selection procedure up. Further methods for the computation of regression for the selection of subsets are given in [Mille 02]. The methods presented there perform in the same order of magnitude in terms of complexity.

Having $G^{\mathcal{S}}$ determined, the actual selection algorithm can be started. A simple algorithm which is suitable for a monotonic measure such as correlation is the *best first search*:

1. Select the best single feature according to $G^{\mathcal{S}}$ for the initialization of $\mathcal{S}$.

2. Evaluate all possible combinations of $\mathcal{S}$ and one additional feature which is not already included in $\mathcal{S}$.

3. Select the best combination to be included in $\mathcal{S}$.

4. If enough features are selected, terminate; else jump to line 2.

For non-monotonic quality criteria, many more complex selection algorithms exist, e.g. *floating forward search* [Niema 03, p.246]. However, they are not required to understand the contents of this work.

Sammon Mapping

A nonlinear method to reduce the dimensionality of a feature space is the Sammon Mapping or the Sammon Transform (ST). It maps high-dimensional data to a plane or a 3-D space [Sammo 69]. In the late 1970's, a fast-converging algorithm for a generalized Sammon transform was presented by Niemann [Niema 79].

Since then, the use of the Sammon transformation has become more and more popular. Especially in the field of data selection and visualization the Sammon maps are often used. Visualizations of the dependencies between different speakers like presented in [Shoza 04] have many fields of application. They can be used to select a subset of representative training speakers in order to reduce the number of training speakers while recognition performance stays in the same range [Nagin 05].

The visualization can also reveal the relations between patients with voice disorders in different graduations [Hader 06b]. Projection of new speakers allows to compare them to the other speakers. This gives a better understanding of the different disorders. Figure 4.9 shows a map of speakers with different degrees of hoarseness. On the top left, speakers with a substitute voice are found. In these patients the larynx was removed due to cancer. The artificial voice of the laryngectomized speakers can be interpreted as an extreme form of hoarseness. The average age of the laryngectomees was about 60 years. At the top right, an age-matched control group of normal speakers is located. At the bottom of the map are speakers with chronic hoarseness.

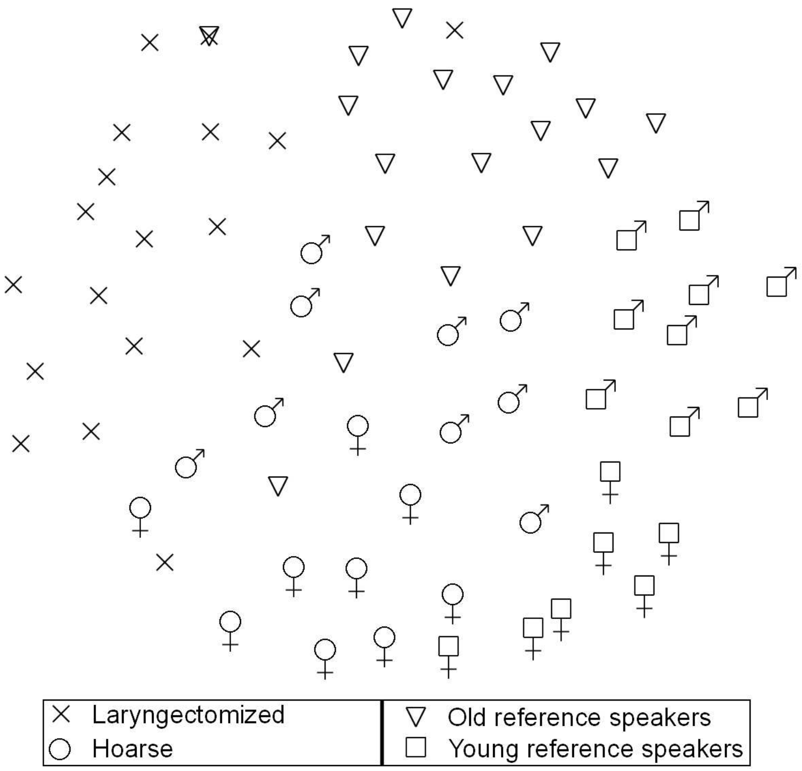


Figure 4.9: Visualization of voice disorders: The properties of the speaker's voices are visible in the map. While the y-axis contains the age of the speaker, the x-axis can be interpreted as the degree of hoarseness of the speakers.

On the bottom right, young reference speakers are found. Hence, the axes of the map can be interpreted as the age on the y-axis and the degree of hoarseness on the x-axis. All data were gathered with the same microphone and with the same recording setup.

As already mentioned, the ST uses the distances between the high-dimensional data to find a low-dimensional representation — called map in the following — that preserves the topology of the original data, i.e. keeps the distances between the low-dimensional representation — called star in the following — as close as possible to the original distances. Doing so, the ST is cluster-preserving. To ensure this, the function ϵ_S is used as a measurement of the error of the resulting map (2-D case):

$$\begin{aligned} \epsilon_\mathrm{S} &= s \sum_{p=1}^{N-1} \sum_{q=p+1}^{N} \frac{(\delta_{pq} - \theta_{pq})^2}{\delta_{pq}} \quad \text{with} \qquad (4.73) \\ \theta_{pq} &= \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2} \qquad (4.74) \end{aligned}$$

δ_{pq} is the high-dimensional distance between the high-dimensional representations of p and q, θ_{pq} is the Euclidian distance between the corresponding stars p and q in the map. s is a scaling factor derived from the high-dimensional distances:

$$s = \frac{1}{\sum_{p=1}^{N-1} \sum_{q=p+1}^{N} \delta_{pq}} \qquad (4.75)$$

The transformation is started with randomly initialized positions for the stars. Then the position of each star is optimized using a conjugate gradient descent library [Naylo 07] and the following gradient:

$$\frac{\partial \epsilon_{\mathrm{S}}}{\partial q_x} = s \sum_{p=1}^{N-1} \sum_{q=p+1}^{N} \frac{2(p_x - q_x)(\delta_{pq} - \theta_{pq})}{\sqrt{(p_x - q_x)^2 + (p_y - q_y)^2}} \tag{4.76}$$

To project a new star q' into an existing map, the high-dimensional distances of the new star and all the stars in the map is needed. The new star is initialized randomly on the map. Then the error

$$\epsilon_{\mathrm{SR}} = s \sum_{p=1}^{N} \frac{(\delta_{pq'} - \theta_{pq'})^2}{\delta_{pq'}} \tag{4.77}$$

is minimized using gradient descent where the original map contained N stars [Exner 07].

Extended Sammon Mapping

For routine clinical use [Noth 07], however, the use of multiple microphones and recording setups is required which poses a serious problem. Modern Internet technologies allow for the recording of speech data at various locations simultaneously in multi-site studies [Maier 07b]. This also means that all data are recorded in different conditions with different microphones. Recording conditions have a great influence. Major factors are the microphone, the distance between the microphone and the speaker, and the acoustical properties of the recording environment. Given a speaker uttering a sentence was recorded simultaneously by multiple microphones of different characteristics at different distances, the points representing the same speaker in the map are spread across the result of the visualization. Figure 4.10 gives an extreme example: The speakers form two clusters although the speakers were recorded simultaneously. This is caused by the acoustic difference between the two microphones which were chosen for the recording. The two corresponding representations of the same speaker are far away from each other in this visualization. The dominating factor is the microphone in this example.

An extension to the standard Sammon transform can be formulated to solve this problem. The extension additionally optimizes the distances between the stars according to information about the data to be mapped. This information could be the age of a certain speaker, his intelligibility, or the membership to a certain group. In our case we assign the same group to the same speaker recorded with multiple microphones.

Now, a grouping error is introduced to extend the objective function.

$$\mathcal{G} = \begin{pmatrix} g_{1,1} & \cdots & g_{1,N} \\ \vdots & \ddots & \vdots \\ g_{N,1} & \cdots & g_{N,N} \end{pmatrix} \tag{4.78}$$

$g_{i,j}$ indicates the distance between the stars, respectively the high-dimensional features, according to the additional information source. Thus, $g_{i,j} = 1$ if the feature vector j is the same as feature vector i, and $g_{i,j} = 0$ if they have nothing in common.

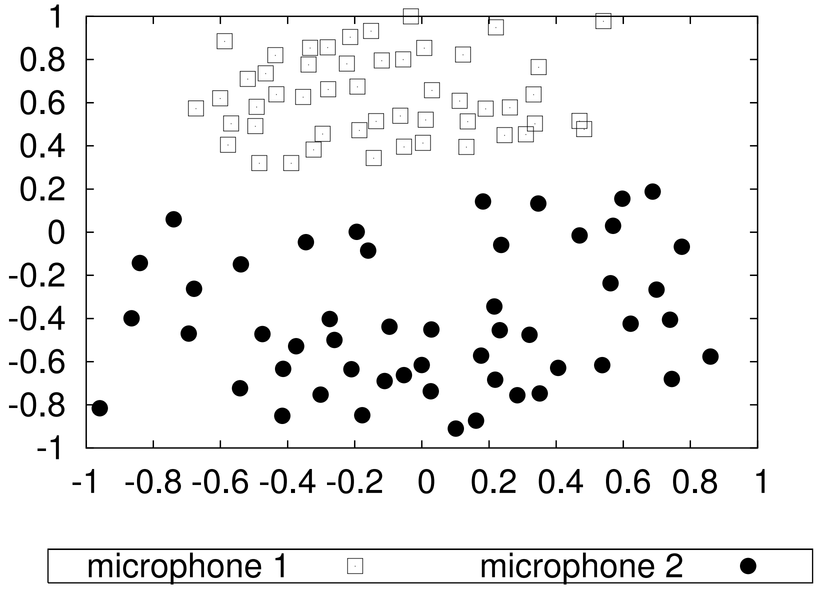


Figure 4.10: 51 speakers recorded simultaneously with two different microphones (remote and close-talk recordings): The two microphones form two clusters although both clusters contain the same speakers [Maier 08c].

The original error function of the Sammon transform is altered such that it penalizes the distance between stars according to their relation in the additional information. A new error function ϵ_{SE} is formed:

$$\epsilon_{\mathrm{SE}} = s \sum_{p=1}^{N-1} \sum_{q=p+1}^{N} \left[w_{\mathrm{S}} g_{p,q} \theta_{pq} + (1 - w_{\mathrm{S}})(1 - g_{p,q}) \frac{(\delta_{pq} - \theta_{pq})^2}{\delta_{pq}} \right] \tag{4.79}$$

$g_{p,q}$ is the group indicator and w_{S} is a weight factor for balancing the standard Sammon error to the additional error term.

Again, gradient descent is applied. Partial derivation leads to the following gradient:

$$\frac{\partial \epsilon_{\mathrm{SE}}}{\partial q_x} = s \sum_{p=1}^{N-1} \sum_{q=p+1}^{N} \left[w_{\mathrm{S}} g_{p,q} \frac{-(p_x - q_x)}{\sqrt{(p_x - q_x)^2 + (p_y - q_y)^2}} + \right. \\ \left. (1 - w_{\mathrm{S}})(1 - g_{p,q}) \frac{2(p_x - q_x)(\delta_{pq} - \theta_{pq})}{\sqrt{(p_x - q_x)^2 + (p_y - q_y)^2}} \right] \tag{4.80}$$

The derivation for the other coordinates is formed analogous.

Choosing the right weight factor w_{S} is crucial to achieve a good reduction of recording conditions influences while keeping the mapping error low, i.e. not to remove the original information presented in the Sammon map. Also the factor must not be chosen too high because it would corrupt the results when projecting new points into an existing map. The re-projection is performed mostly in the same way as for the standard ST since the group information of a new star will be known a priori [Maier 08c].

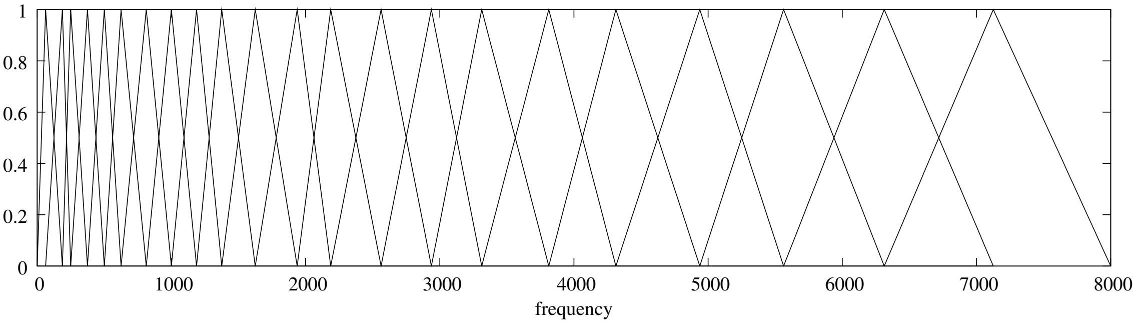


Figure 4.11: The composition of the Mel filter bank with 22 filters

4.2.4 Speech Processing

Automatic speech processing concerns all topics which are required to perform speech recognition. In order to give a short overview on this field, first feature extraction is described. The standard features in speech recognition are *Mel Frequency Cepstrum Coefficients* (MFCCs). One of the possible fields of application of MFCCs is speaker recognition with *Gaussian Mixture Models* (GMMs). As mentioned before the GMMs are crucial in order to separate children's speech from adults' speech. Next, the actual speech recognizer is presented. It consists of acoustic models — the so-called *Hidden Markov Models* (HMMs) and a statistical *Language Model*. After the examination of the training procedures for both models, their joint decoding can be performed: The actual speech recognition process. A short history of speech recognition is given in [Furui 05].

In addition to speech recognition, two more points are explained in this section: Prosodic analyses which extract information on the speaking and intonation style and pronunciation analysis which models common pronunciation errors.

Feature Extraction

We assume that the signal $f(t)$ is sampled in equidistant steps and quantized at 16 bit. Usually the sampling rate varies from 8 kHz (telephone) to 44.1 kHz (CD). In our experiments, all data were sampled with 16 kHz. Since the signal is an audio stream, it is one-dimensional. Now feature vectors $\boldsymbol{x}(t)$ are extracted from the signal at each discrete time t. If the dimension of the feature vector is n, it can be denoted as

$$\boldsymbol{x}(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{pmatrix} = \boldsymbol{x}_t \ . \tag{4.81}$$

In the following the extraction of MFCCs which are being used for more than 20 years in speech recognition is presented [Davis 80]. They are obtained from the spectrum of the speech signal [Niema 03]. The spectrum can be computed from the

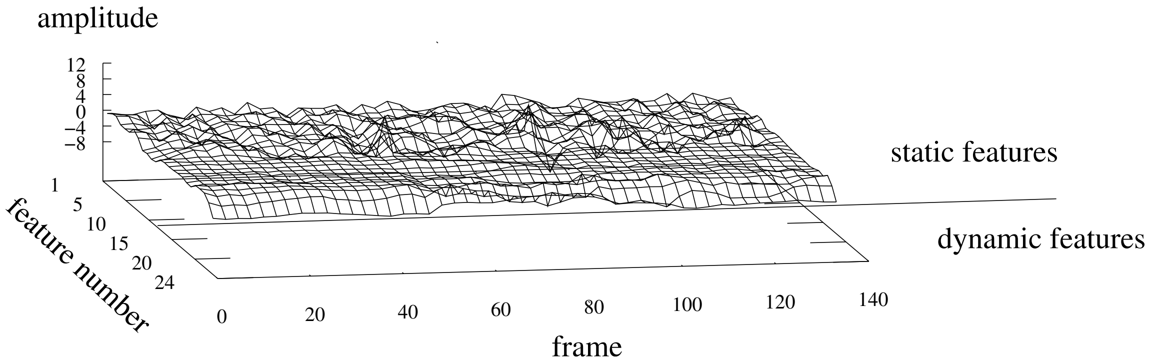


Figure 4.12: MFCC features as a 3-D plot: Note that the range of the static features is much wider than the range of the dynamic features.

speech signal by Fourier analysis. Since we assume to process sampled and quantized data, only the discrete Fourier transform will be introduced here:

$$\begin{aligned} \tilde{F}_\mu &= \sum_{j=0}^{I-1} \tilde{f}_j e^{-i2\pi\left(\frac{\mu j}{I}\right)} &&= DFT\{\tilde{f}_j\} \\ \tilde{f}_j &= \frac{1}{I}\sum_{\mu=0}^{I-1} \tilde{F}_\mu e^{i2\pi\left(\frac{\mu j}{I}\right)} &&= DFT^{-1}\{\tilde{F}_\mu\} \end{aligned} \tag{4.82}$$

I denotes the length of the signal to be transformed while $\tilde{f}_j$ and $\tilde{F}_\mu$ stand for the periodic continuation of the signal ($\tilde{f}_{I+1} := f_1$) and its Fourier transformation. The correctness can easily be shown with $\tilde{f}_j = DTF^{-1}\{DFT\{\tilde{f}_j\}\}$ [Niema 03].

In order to analyze the change of the frequencies over time, the signal is split to short chunks (frames) which are transformed separately. If a rectangular splitting window is applied, high frequencies can occur because the signal is cut off sharply at the edges of the time frame. This can be attenuated by window functions like the Hamming or the Hanning window [Niema 03, Maier 05a, p.9]. In our case this analysis is done for a period of 16 ms with a shift of 10 ms. The overlap is done to compensate the loss of information caused by the window functions. A spectrogram showing the progression of the frequencies can be generated.

A lot of the information included in the spectrum is not audible by the human ear [Moore 86]. Hence, the amount of data can be reduced at this point [Zwick 67]. This is usually done by a filter bank. For the case of MFCCs, 22 filters were used. Figure 4.11 plots the triangular filters. Note that the lower frequencies are analyzed better — since more filters are computed and their range is smaller — than the higher frequencies, just like by the human ear. In each of the filters, the energy is integrated.

In the next step, the base 10 logarithm is applied to the 22 values obtained by the filter bank. Again, this is done to resemble the characteristics of the human ear. Plotted in a spectrographic layout, this is known as Mel spectrogram.

In order to compute cepstral parameters, the logarithmic Mel spectrum is inverted using an inverse discrete cosine transform (iDCT). Note that the word **ceps**trum is an inverse — of the word **spec**trum — as well. The Mel cepstrum is formed. In this thesis the first 12 cepstral coefficients (c_{0-11}) are used for feature computation since the low coefficients are more important for speech recognition. However, the first magnitude (c_0) is replaced by the log energy because it contains less important information than the energy. Furthermore, the first derivative for these values is computed using a regression line over 5 consecutive frames which yields additional 12 coefficients [Furui 00]. In Figure 4.12 a 3-D plot of MFCC features can be seen. The amplitude sequence of the dynamic features is much smoother than the amplitudes of the static features. The range of the dynamic features is smaller.

Speaker Recognition

The most common method to perform speaker and speaker group recognition are *Gaussian Mixture Models* (GMMs, [Hertl 99, Cinca 02, Bockl 07b, Bockl 07c]). However, different approaches exist as well. Overviews are found in [Furui 91, Furui 94, Furui 97]. In general, the task of speaker recognition is to determine the speaker S of an utterance $\boldsymbol{X}$. Therefore, the properties of all N_S speakers have to be modeled with a GMM $\boldsymbol{\lambda}_{\mathrm{GMM}}$. Classification is then performed according to Bayes' rules by the selection of the $\boldsymbol{\lambda}_{\mathrm{GMM}}$ with the highest a posteriori probability

$$P(\boldsymbol{\lambda}_{\mathrm{GMM}} \mid x) = \frac{p(x \mid \boldsymbol{\lambda}_{\mathrm{GMM}}) P(\boldsymbol{\lambda}_{\mathrm{GMM}})}{p(x)}. \tag{4.83}$$

Each GMM is composed of M *Gaussian distributions.* Each distribution is weighted by $w_m \in (0,1)$ with $\sum_{m=1}^{M} w_m = 1$. Each speaker is modeled by M n-dimensional weighted densities:

$$\begin{aligned} p(\boldsymbol{x} \mid \boldsymbol{\lambda}_{\mathrm{GMM}}) &= \sum_{m=1}^{M} w_m \mathcal{N}_m(\boldsymbol{x}) \quad \text{with} & (4.84) \\ \mathcal{N}_m(\boldsymbol{x}) &= \frac{1}{(2\pi)^{n/2} \mid \boldsymbol{\Sigma}_m \mid^{1/2}} e^{-(1/2)(\boldsymbol{x}-\boldsymbol{\mu}_m)^T \boldsymbol{\Sigma}_m^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_m)} & (4.85) \end{aligned}$$

Each of the GMMs $\boldsymbol{\lambda}_{\mathrm{GMM}}$ is determined by the parameters $w_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m$, where $\boldsymbol{\mu}_m$ is the mean value, $\boldsymbol{\Sigma}_m$ the covariance matrix, and w_m the weight of distribution m.

Figure 4.13 shows the first two dimensions of the GMMs of three different speakers. Each speaker is modeled with four Gaussian distributions. Each distribution characterizes different acoustic properties of the speakers. Although all distributions of the speakers are different, the positions of the mean vectors and the covariances resemble each other.

Next, these maximum-likelihood parameters have to be estimated on a given training set. The *Expectation-Maximization* (EM) algorithm [Demps 77, Redne 84] can be applied to solve the problem of the model parameter estimation with a given training set. The algorithm iteratively refines the GMM parameters by maximizing the likelihood of an estimated model using the observed training feature vectors $\boldsymbol{X}$. In each iteration a new model $\tilde{\boldsymbol{\lambda}}_{\mathrm{GMM}}$ with increased likelihood is estimated from an initial model $\boldsymbol{\lambda}_{\mathrm{GMM}}$ according to:

$$p(\boldsymbol{X} \mid \tilde{\boldsymbol{\lambda}}_{\mathrm{GMM}}) \geq p(\boldsymbol{X} \mid \boldsymbol{\lambda}_{\mathrm{GMM}}) \tag{4.86}$$

The EM algorithm consists of the following four steps:

- Determination of initial model parameters
- E step: Calculate a posteriori probabilities

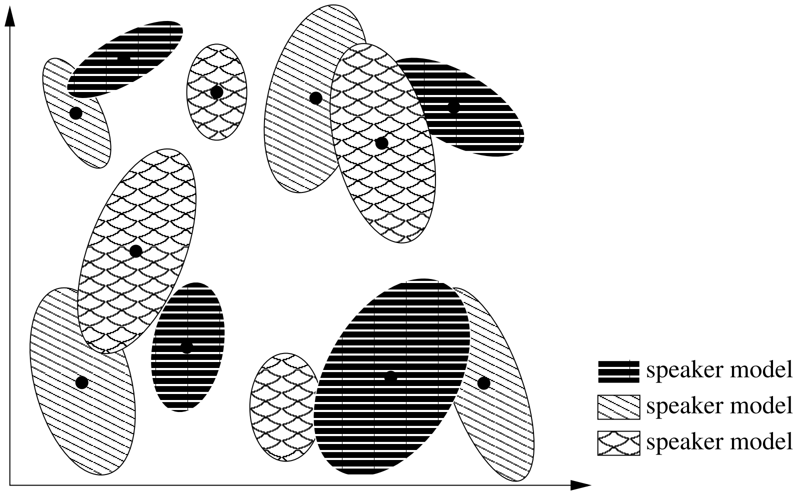


Figure 4.13: 2-D example of speaker models each using 4 Gaussian mixtures — after [Bockl 07b]

- M step: Calculate new estimation values of the maximum-likelihood parameters $w_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m$
- repeat E and M step until convergence

The initialization is performed by a vector quantization (VQ) algorithm. It creates an initial segmentation with M densities from the given training set. This initial segmentation is then used to determine a first parameter set. Next, these parameters are improved iteratively with the EM algorithm [Niema 03]. The most commonly used VQ algorithms are the k-means [Ander 73] and the LBG approach [Linde 80, Buzo 80], named after the inventors Linde, Buzo and Gray.

In the E step, the a posteriori probabilities of the feature vectors $\boldsymbol{x}_t$ for every mixture m of each GMM $\boldsymbol{\lambda}_{\text{GMM}}$ are calculated by a modification of the Bayesian formula Eq. 4.83

$$p(m \mid \boldsymbol{x}_t, \boldsymbol{\lambda}_{\text{GMM}}) = \frac{w_m \mathcal{N}_m(\boldsymbol{x}_t)}{\sum_{i=1}^{M} w_i \mathcal{N}_i(\boldsymbol{x}_t)}. \tag{4.87}$$

With these posterior probabilities, the parameters of the improved model are then reestimated in the M step:

$$\tilde{w_m} = \frac{1}{T} \sum_{t=1}^{T} p(m \mid \boldsymbol{x}_t, \boldsymbol{\lambda}_{\text{GMM}}) \tag{4.88}$$

$$\tilde{\boldsymbol{\mu}}_m = \frac{\sum_{t=1}^{T} p(m \mid \boldsymbol{x}_t, \boldsymbol{\lambda}_{\text{GMM}}) \boldsymbol{x}_t}{\sum_{t=1}^{T} p(m \mid \boldsymbol{x}_t)} \tag{4.89}$$

$$\tilde{\boldsymbol{\Sigma}}_m = \frac{\sum_{t=1}^{T} p(m \mid \boldsymbol{x}_t, \boldsymbol{\lambda}_{\text{GMM}})}{\sum_{t=1}^{T} p(m \mid \boldsymbol{x}_t)} (\boldsymbol{x}_t - \tilde{\boldsymbol{\mu}_m})(\boldsymbol{x}_t - \tilde{\boldsymbol{\mu}_m})^\top \tag{4.90}$$

After the M step, the model $\boldsymbol{\lambda}_{\text{GMM}}$ is set to the new model $\tilde{\boldsymbol{\lambda}}_{\text{GMM}}$. These two steps are continued until convergence, or a fixed number of iterations has been reached.

Since training data is always sparse, the use of a *Universal Background Model* (UBM) showed to improve the results of speaker recognition in the literature [Reyno 95, Reyno 00, Yang 04, Brand 05, Deng 05]. The UBM is usually trained on all available training data to represent as much features of speech as possible. Then *Maximum A Posteriori* (MAP) adaptation [Gauva 94] is used to derive speaker-adapted models from the well-trained UBM parameters with the speaker's training data. The models created from a UBM are called *coupled models.*

MAP adaptation for Gaussian Mixtures consists of two steps, like the EM algorithm described above. The first step determines the a posteriori probabilities with the UBM parameters. In the second step, w_m, $\boldsymbol{\mu}_m$, and $\boldsymbol{\Sigma}_m$ are estimated from each speaker's data. Given T training vectors of the speakers $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T\}$ as training set and a UBM $\boldsymbol{\lambda}_{\mathrm{UBM}}$, the a posteriori probability $P(m \mid \boldsymbol{x}_t)$ of mixture m given feature vector $\boldsymbol{x}_t$ is determined as

$$P(m \mid \boldsymbol{x}_t) = \frac{w_m \mathcal{N}_m(\boldsymbol{x}_t)}{\sum_{i=1}^{M} w_i \mathcal{N}_i(\boldsymbol{x}_t)} \tag{4.91}$$

where $\mathcal{N}_m(\boldsymbol{x}_t)$ is the Gaussian distribution from Eq. 4.85. Using $P(m \mid \boldsymbol{x}_t)$ and the feature vectors $\boldsymbol{x}_t$, ML estimates of the weight w'_m, the mean $\boldsymbol{\mu}'_m$, and the variance $\boldsymbol{\Sigma}'{}_m$ of each mixture m are calculated.

$$w'_m = \sum_{t=1}^{T} P(m \mid \boldsymbol{x}_t) \tag{4.92}$$

$$\boldsymbol{\mu}'_m = \sum_{t=1}^{T} P(m \mid \boldsymbol{x}_t)\boldsymbol{x}_t \tag{4.93}$$

$$\boldsymbol{\Sigma}'_m = \sum_{t=1}^{T} P(m \mid \boldsymbol{x}_t)\boldsymbol{x}_t\boldsymbol{x}_t^{\top} \tag{4.94}$$

In order to create an adapted GMM, these ML estimates can be combined with the UBM parameter linearly:

$$\hat{w_m} = [\alpha'_m w_m / T + (1 - \alpha'_m) w_m]\gamma \tag{4.95}$$

$$\hat{\boldsymbol{\mu}_m} = \alpha'_m \boldsymbol{\mu}'_m + (1 - \alpha'_m)\boldsymbol{\mu}_m \tag{4.96}$$

$$\hat{\boldsymbol{\Sigma}}_m = \alpha'_m \boldsymbol{\Sigma}'_m + (1 - \alpha'_m)(\boldsymbol{\Sigma}_m + \boldsymbol{\mu}_m \boldsymbol{\mu}_m^{\top}) - \boldsymbol{\mu}_m \boldsymbol{\mu}_m^{\top} \tag{4.97}$$

where α'_m is an adaptation parameter for the linear combination. α'_m is dependent on the size of the adaptation data N_m

$$\alpha'_m = \frac{N_m}{N_m + \tau}, \tag{4.98}$$

and the relevance parameter τ which has to be defined by the user. Is the number of feature vectors of a given density N_m sparse, then $\alpha'_m \to 0$. Thus densities which were observed seldom cause only small changes in the adaptation parameters. On the contrary, a high count ($\alpha'_m \to 1$) leads to high variations in the adapted parameters.

Acoustic Modeling for Speech Recognition

A speech recognizer finds the most likely word chain $w_1^*, \ldots, w_S^*$ for a given sequence of observations $\boldsymbol{o}^T := (o_1, o_2, \ldots, o_T)$, $o_t \in \Omega$, $t \in [1, \ldots, T]$, of length T. Acoustic modeling is used to model the acoustic properties of a certain word or phoneme in order to recognize them. The state-of-the-art approach in speech recognition are *Hidden Markov Models* (HMMs) $\boldsymbol{\lambda}_{\text{HMM}}$. These models are able to compute the probability $P_{\text{AM}}(\boldsymbol{o}^T \mid \boldsymbol{w}^S)$ of a given sequence of T observations $\boldsymbol{o}^T$ [Huang 01]. Each o_t is element of one class Ω_κ, $\kappa \in [1, \ldots, K]$, where K denotes the total number of classes. The probability for a certain observation $p(\boldsymbol{o}^T)$ can be written as the product of conditional probabilities:

$$p(\boldsymbol{o}^T) = p(o_1) \prod_{t=2}^{T} p(o_t \mid o_1, \ldots, o_{t-1}) \tag{4.99}$$

The Markov assumption that the next observation is only dependent on the current one can reduce this product to:

$$p(\boldsymbol{o}^T) = p(o_1) \prod_{t=2}^{T} p(o_t \mid o_{t-1}) \tag{4.100}$$

Now we define that the model $\boldsymbol{\lambda}_{\text{HMM}}$, which models the word chain $\boldsymbol{w}^S$, has N_s hidden states s_i. Hidden states can emit any of the observations o_t at any discrete time step t. However, the sequence of the states is unknown to the observer. The probability to emit a certain observation o_t differs in every hidden state. Thus, a hidden state variable q_t is introduced.

$$p(\boldsymbol{o}^T \mid \boldsymbol{\lambda}_{\text{HMM}}) \approx \sum_{\forall q_t \in \{s_1, \ldots, s_{N_s}\} \forall t=1,\ldots,T} P(q_1)\, p(o_1 \mid q_1) \prod_{t=2}^{T} p(o_t \mid q_t)\, P(q_t \mid q_{t-1}) \tag{4.101}$$

$P(q_1 = s_i)$ denotes the probability to start in state s_i, and $P(q_t = s_j \mid q_{t-1} = s_i)$ describes the transition probability from state s_i at time $t-1$ to state s_j at time t. Hence, λ_{HMM} can be defined as a triplet $(\boldsymbol{\pi}, \mathfrak{A}, \mathfrak{B})$ where $\boldsymbol{\pi}$ is a vector of size N_s which contains the start probabilities $P(q_1 = s_i)$, $\mathfrak{A}$ is an $N_s \times N_s$ matrix with the transition probabilities $P(q_t = s_j \mid q_{t-1} = s_i)$, and $\mathfrak{B}$ is another $N_s \times K$ matrix with the output probabilities for each observation $P(o_t \in \Omega_\kappa \mid q_t)$ in every state.

Until now, no restrictions for the transition were done. For example it is possible to jump from any state to any other state. In continuous speech recognition, this case is highly unlikely. We assume that transitions only happen to the next state in positive time direction which is the case in normal speech[3]. In addition, transitions to the current state are allowed, too, in order to model variations in speaking rate. Thus the transition matrix $\mathfrak{A}$ can be reduced to a $N_s \times 2$ matrix. This type of HMMs is called *linear HMM* [Schuk 95]. With this definition of HMMs, several problems can be stated and solved:

[3] Stuttering, for example, can be modeled by state transitions in negative time direction [Noth 00].

- **Forward Algorithm:** The most likely model $\boldsymbol{\lambda}^*_{\mathrm{HMM}}$ for an observation $\boldsymbol{o}^T$ can be created by choosing the model $\boldsymbol{\lambda}_{\mathrm{HMM}}$ which generates the highest probability $P(\boldsymbol{o}^T \mid \boldsymbol{\lambda}_{\mathrm{HMM}})$ for a given sequence of observations.

$$\boldsymbol{\lambda}^*_{\mathrm{HMM}} := \underset{\boldsymbol{\lambda}_{\mathrm{HMM}}}{\operatorname{argmax}} P(\boldsymbol{o}^T \mid \boldsymbol{\lambda}_{\mathrm{HMM}}) \tag{4.102}$$

Hence, the problem can be reduced to computing $P(\boldsymbol{o}^T \mid \boldsymbol{\lambda}_{\mathrm{HMM}})$ for each model $\boldsymbol{\lambda}_{\mathrm{HMM}}$. This can be done with the so-called *forward algorithm*. According to [Schuk 95] the forward probability $\alpha_t(j)$ can be computed recursively as

$$\begin{aligned}\alpha_t(j) &= P(\boldsymbol{o}^T, q_t = j \mid \boldsymbol{\lambda}_{\mathrm{HMM}}) \\ &= \left(\sum_{i=1}^{N_s} \alpha_{t-1}(i) a_{ij}\right) b_j(o_t) \qquad \text{with } \alpha_1(j) = \pi_j b_j(o_1)\ . \end{aligned} \tag{4.103}$$

The matrix elements π_j, a_{ij} and b_{jk} can be obtained from the corresponding matrices $\boldsymbol{\pi}$, $\mathfrak{A}$, and $\mathfrak{B}$. Using the forward probability $\alpha_t(j)$, the probability $P(\boldsymbol{o}^T \mid \boldsymbol{\lambda}_{\mathrm{HMM}})$ can now be found by

$$P(\boldsymbol{o}^T \mid \boldsymbol{\lambda}_{\mathrm{HMM}}) = \sum_{j=1}^{N_s} \alpha_t(j)\ . \tag{4.104}$$

- **Viterbi Algorithm:** The *Viterbi algorithm* can compute the most likely sequence of states $\boldsymbol{q}^{T*}$. Hence, the probability $P(\boldsymbol{q}^T \mid \boldsymbol{o}^T, \boldsymbol{\lambda}_{\mathrm{HMM}})$ has to be maximized. In [Schuk 95] it is shown that this can be done by a recursive calculation of the maximal probability $P^*(\boldsymbol{o}^T \mid \boldsymbol{\lambda}_{\mathrm{HMM}})$ which can be reached for a sequence of observations $\boldsymbol{o}^T$, given a certain model $\boldsymbol{\lambda}_{\mathrm{HMM}}$. This is done by computation of the maximum probabilities $\theta_t(j)$ and a matrix which stores the best previous state $\phi_t(j)$ at time t for state s_j simultaneously. The recursion is initialized with

$$\forall_j : \quad \theta_1(j) = \pi_j b_j(o_1) \qquad \text{and} \qquad \forall_j : \quad \phi_1(j) = 0. \tag{4.105}$$

Then new values are computed in each recursion step by

$$\forall_j : \quad \theta_t(j) = \max_i \left(\theta_{t-1}(j) a_{ij}\right) b_j(o_t) \quad \text{and} \quad \forall_j : \quad \phi_t(j) = \underset{i}{\operatorname{argmax}}\, \theta_{t-1}(i) a_{ij}\ . \tag{4.106}$$

Now $\boldsymbol{q}^{T*}$ can be found as

$$\begin{aligned}\boldsymbol{q}^{T*} &= (q_1^*, q_2^*, \ldots, q_T^*) \qquad \text{with} \\ q_T^* &= \underset{j}{\operatorname{argmax}}\, \theta_t(j) \qquad \text{and} \qquad q_t^* = \phi_t(q_{t+1}^*)\ .\end{aligned} \tag{4.107}$$

In this manner the most likely sequence of states can be found.

- **Baum-Welch Algorithm:** In order to determine the parameter matrices $\boldsymbol{\pi}$, $\mathfrak{A}$, and $\mathfrak{B}$, the Markov models have to be trained with $\boldsymbol{o}^T$ which is now regarded as the training set. It is used in order to estimate the model parameters. This is done with the *Baum-Welch algorithm.* The forward probabilities $\alpha_t(j)$ which were already given in Eq. 4.103 are applied again. Furthermore, backward probabilities $\beta_t(j)$ are needed as well. In [Schuk 95] they are defined as

$$\beta_t(i) = \sum_{i=1}^{N_s} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j) \qquad \text{with} \quad \beta_T(i) = 1 \; . \tag{4.108}$$

Since the forward and the backward probabilities are needed, the algorithm is also known as *forward-backward algorithm.* Furthermore, the probability $\gamma_t(j)$ that the hidden state q_t is state s_j at time t is needed, too.

$$P(q_t = s_j | \boldsymbol{O}, \boldsymbol{\lambda}_{\text{HMM}}) = \frac{P(\boldsymbol{O}, q_t = j | \boldsymbol{\lambda}_{\text{HMM}})}{P(\boldsymbol{O} | \boldsymbol{\lambda}_{\text{HMM}})} = \frac{\alpha_t(j)\beta_t(j)}{\sum_i \alpha_t(i)\beta_t(i)} = \gamma_t(j) \tag{4.109}$$

Now the maximum likelihood estimates of the parameters of $\boldsymbol{\lambda}$ are computed by

$$\hat{\pi}_i = \gamma_1(i) \tag{4.110}$$

$$\hat{a}_{ij} = \frac{\sum\limits_t \alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\sum\limits_t \alpha_t(i)\beta_t(i)} \tag{4.111}$$

$$\hat{b}_{j\kappa} = \frac{\sum\limits_t \gamma_t(j) \chi_{[o_t \in \Omega_\kappa]}}{\sum\limits_t \gamma_t(j)} . \tag{4.112}$$

$\chi_{[o_t \in \Omega_\kappa]}$ is the characteristic function which returns 1 if o_t is element of Ω_κ and 0 if it is not. In this manner all model parameters can be estimated. However, the o_t are discrete symbols for one class Ω_κ. Hence, every feature vector $\boldsymbol{x}_t$ has to be classified first. It can be assigned to a certain class Ω_κ. These kind of HMMs are known as *discrete hidden Markov models.*

With *semi-continuous hidden Markov models* (SCHMM), the hard classification of one vector to exactly one class as assumed before is smoothened. The features are *soft-quantized* [Schuk 95]. Therefore, a codebook with M Gaussian densities is introduced. For each density m, the mean $\boldsymbol{\mu}_m$ and the covariance $\boldsymbol{\Sigma}_m$ are computed. Probabilities for multidimensional feature vectors can be calculated. Unlike *continuous HMMs*, which have one codebook per state, SCHMMs share one codebook for all states and models.

The $b_j(o_t)$ are now computed for feature vectors $\boldsymbol{x}_t$ by

$$b_j(\boldsymbol{x}_t) = \sum_{m=1}^{M} c_{jm} \mathcal{N}_m(\boldsymbol{x}_t), \qquad \sum_{m=1}^{M} c_{jm} = 1 \; . \tag{4.113}$$

$\mathcal{N}_m(\boldsymbol{x}_t)$ is a unimodal Gaussian distribution according to Eq. 4.85. By mixture of a sufficient number of components, any density function can be approximated.

Again, maximum likelihood parameters for $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$ can be estimated using a training set $\boldsymbol{x}^T = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T)$. To do this the probability $\zeta_t(j,m)$ to select component m at time t in state s_j is needed. It can be calculated with the forward (Eq. 4.103) and backward (Eq. 4.108) probabilities by

$$\begin{aligned} \zeta_t(j,m) &= P(q_t = s_j, m_t = m \mid \boldsymbol{x}^T, \boldsymbol{\lambda}_{\mathrm{HMM}}) && (4.114) \\ &= \begin{cases} \dfrac{1}{P(\boldsymbol{x}^T \mid \boldsymbol{\lambda}_{\mathrm{HMM}})} \displaystyle\sum_i \alpha_{t-1}(i)\, a_{ij}\, c_{jm} \mathcal{N}_m(\boldsymbol{x}_t) \beta_t(j) & t \geq 1 \\ \dfrac{1}{P(\boldsymbol{x}^T \mid \boldsymbol{\lambda}_{\mathrm{HMM}})} \displaystyle\sum_i \pi_j\, c_{jm} \mathcal{N}_m(\boldsymbol{x}_t) \beta_1(j) & t = 1 \end{cases} . \end{aligned}$$

Now the estimates can be computed as

$$\hat{c}_{jm} = \frac{1}{\displaystyle\sum_t \gamma_t(j)} \sum_t \zeta_t(j,m) \tag{4.115}$$

$$\hat{\boldsymbol{\mu}}_m = \frac{1}{\displaystyle\sum_t \zeta_t(j,m)} \sum_t \zeta_t(j,m) \boldsymbol{x}_t \tag{4.116}$$

$$\hat{\boldsymbol{\Sigma}}_m = \frac{1}{\displaystyle\sum_t \zeta_t(j,m)} \sum_t \zeta_t(j,m) \boldsymbol{x}_t \boldsymbol{x}_t^\top - \boldsymbol{\mu}_m \boldsymbol{\mu}_m^\top . \tag{4.117}$$

Next, both the codebook and the Markov models can be trained. Both estimations are done in an alternating manner because new HMM parameters will produce new posteriors $\zeta_t(j,m)$ (cf. Eqs. 4.115, 4.116, and 4.117) while new codebook densities refine the values of $b_j(\boldsymbol{x}_t)$ (cf. Eq. 4.111). First an initial codebook is estimated by identification of a Gaussian mixture distribution with M components with the training set $\boldsymbol{x}^T$. In addition, initial Markov models with uniform transition probabilities are built. Then the re-estimation of the models, using Baum-Welch training, and the codebook starts. In this work 10 re-estimations of the models are done followed by one re-estimation of the codebook in each re-estimation step. During the training, 10 such re-estimation steps are done in total.

Language Modeling

With a language model, further linguistic information is supplied to the recognition process. This is accomplished with various approaches using formal grammars or stochastic information. While formal grammars are useful to create quite restricted speech recognition systems, like a phone-number recognition system or a simple voice command system, stochastic language models are state-of-the-art in spontaneous speech recognition systems. Therefore, only stochastic language models are presented at this point.

So-called *n-gram* models are used to describe the linguistic information. These n-grams give the probability for a certain word w_i in a context of $n-1$ preceding words. For a word sequence $\boldsymbol{w}^S$, the probability $P_{\mathrm{LM}}(\boldsymbol{w}^S)$ is approximated as

$$\begin{aligned} P_{\mathrm{LM}}(\boldsymbol{w}^S) &:= P(w_1 \dots w_S) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1 w_2) \dots P(w_S \mid w_1 \dots w_{S-1}) \\ &= P(w_1) \prod_{j=2}^{S} P(w_j \mid w_1, \dots, w_{j-1}) \\ &\approx P(w_1) \dots P(w_{n-1} \mid w_1, \dots, w_{n-2}) \prod_{j=n}^{S} P(w_j \mid w_{j-n+1}, \dots, w_{j-1}) \quad (4.118) \end{aligned}$$

where $n \geq 3$. Further common n-grams are the zerogram model ($n = 0$) where all words w_i are distributed equally, the unigram $P(w_i)$ ($n = 1$), the bigram model $P(w_i \mid w_{i-1})$ ($n = 2$), and the trigram model $P(w_i \mid w_{i-2}, w_{i-1})$ ($n = 3$).

For small values of n, the models are estimated by counting the occurrences of the word w_i in the desired context in a large text corpus. A probability for every word and context is found. Nevertheless, it happens that a certain word in a certain context is not a member of the training corpus but of the test set. For this case the language model would prohibit its recognition since it was never observed. The problem is solved by interpolation: Only words are counted which appear twice or more in a certain context. The probability mass which was not assigned can be distributed among the words and contexts which appeared less than twice. This kind of smoothing is also known as *Good-Turing-Smoothing* (cf. [Nadas 85] and [Schuk 95, p.217]).

If bigger contexts are chosen, like in the case of 4-grams, this simple interpolation is not sufficient since only few of the possible 4-grams are observed even in a large training corpus [Allis 06]. Therefore, categories C_i are employed which are assigned to every word w_i. $P_{\mathrm{LM}}(\boldsymbol{w}^S)$ for n-grams can be written as

$$P(w_i \mid w_{i-n+1}, \dots, w_{i-1}) := P(w_i \mid C_i) \; P(C_i \mid C_{i-n+1}, \dots, C_{i-1}) \qquad (4.119)$$

In this manner the same n-gram context can be observed much more often. Usually the categories are found by linguistic categories like "verbs" and "articles" or semantic categories like "place names" or "player names". Of course, data-driven approaches exist as well [Knese 93].

In this thesis all recognition results are done with a category-based unigram or 4-gram model as already used in [Gallw 02] and [Hader 04]. The categories were chosen semantically to match the words of the test (cf. Chapter 5.1).

	recognized word chain	reference	%
WA	this is moon bucket and a a ball	this is a moon a bucket and a tree	56
WR	this is moon bucket and a a ball	this is a moon a bucket and a tree	67
WA	tiger moon bucket apple ball	moon bucket tree	0
WR	tiger moon bucket apple ball	moon bucket tree	67

Table 4.1: Example of the effects of the automatic reference on the WA and WR. We assume that the spoken utterance is “This is a moon, a bucket, and a tree”. Thus, the automatic reference is “moon bucket tree”.

Decoding

Now that the acoustic model probability $P_{\mathrm{AM}}(\boldsymbol{o}^T \mid \boldsymbol{w}^S)$ and the language model probability $P_{\mathrm{LM}}(\boldsymbol{w}^S)$ are found, the most likely word chain $w_1^*, \ldots, w_S^*$ can be computed by decoding. According to [Stemm 05] this can be written as

$$\begin{aligned} w_1^*, \ldots, w_S^* &:= \operatorname*{argmax}_{\boldsymbol{w}^S} P_{\mathrm{AM,LM}}(\boldsymbol{w}^S \mid \boldsymbol{o}^T) && (4.120) \\ &= \operatorname*{argmax}_{\boldsymbol{w}^S} \frac{P_{\mathrm{AM}}(\boldsymbol{o}^T \mid \boldsymbol{w}^S) P_{\mathrm{LM}}(\boldsymbol{w}^S)}{P(\boldsymbol{o}^T)} \\ &= \operatorname*{argmax}_{\boldsymbol{w}^S} P_{\mathrm{AM}}(\boldsymbol{o}^T \mid \boldsymbol{w}^S) P_{\mathrm{LM}}(\boldsymbol{w}^S). && (4.121) \end{aligned}$$

Unfortunately, Eq. 4.121 can only be applied as is in the case of a zero- or unigram model. For the case of bigger contexts, a search tree has to be built. In order to prune the search tree, a beam search is done. This kind of search filters out non-probable candidates which are outside the search beam. When the tree is built, its branches can be re-scored. In this phase a language model with a large context is applied since the possible context is fixed by the beam search. Then the best word chain is found by the application of the A^* algorithm [Niema 03]. The latest version of the decoding as applied in this work is described in [Stemm 05].

For the evaluation of the decoded sequence, two measures are commonly used: the word accuracy (WA) and the word recognition rate (WR).

$$\mathrm{WR} = \frac{C}{R} \times 100\,\%$$

is computed as the percentage of correctly recognized words C and the number of reference words R. In addition,

$$\mathrm{WA} = \frac{C - I}{R} \times 100\,\%$$

weighs the number of wrongly inserted words I in this percentage. The WA punishes the insertion of additional words compared to the reference chain. The upper limit of both measures is 100 %. The lower bound of the WR is 0 % while the WA does not have a lower bound. It gets negative as soon as the recognizer inserts more wrong words than it actually recognizes correctly. Table 4.1 gives an example for the difference in computation.

Word level features	
Feature	Description
PauseSilenceBeforeWord	Length of the pause before the current word
PauseSilenceAfterWord	Length of the pause after the current word
EnergyRegCoeffWord	Slope of the regression line of the energy contour
EnergyMseRegWord	Mean square error of the regression line of the energy contour
EnergyEneAbsWord	Absolute energy of the current word
EnergyMaxPosWord	Position of the maximal energy in the current word
EnergyMaxWord	Value of the maximal energy in the current word
EnergyMeanWord	Mean value of the energy in the current word
DurLenAbsWord	Duration of the current word
DurLenAbsSyllableWord	Mean duration of the syllables in the current word
F0RegCoeffWord	Slope of the regression line of the F_0 contour in the current word
F0MseRegWord	Mean square error of the regression of the F_0 contour in the current word
F0MaxWord	Maximal F_0 value in the current word
F0MinWord	Minimal F_0 value in the current word
F0MeanWord	Average F_0 value of the current word
F0OnsetWord	First value of the F_0 contour in the current word
F0OffsetWord	Last value of the F_0 contour in the current word
F0OnsetPosWord	Position of the F_0 Onset in the current word
F0OffsetPosWord	Position of the F_0 Offset in the current word
F0MinPosWord	Position of the minimal F_0 value in the current word
F0MaxPosWord	Position of the maximal F_0 value in the current word

Table 4.2: Overview on the prosodic features computed on word level

Prosodic Analysis

The prosody module takes the output of our word recognition module in addition to the speech signal as input. In this case the time-alignment with the Viterbi algorithm of the recognizer and the information about the underlying phoneme classes (like *long vowel*) can be used by the prosody module [Batli 00].

First, the prosody module extracts so-called base features from the speech signal. These are the energy, the fundamental frequency (F_0) after [Bagsh 93], and the voiced and unvoiced segments of the signal. In a second step, the actual prosodic features are computed to model the prosodic properties of the speech signal. Therefore, a fixed reference point has to be chosen for the computation of the prosodic features. We decided in favor of the end of a word because the word is a well–defined unit in word recognition, it can be provided by any standard word recognizer, and because this point can be more easily defined than, for example, the middle of the syllable nucleus in word accent position. For each reference point, we extract 21 prosodic features (cf. Table 4.2). These features model F_0, energy and duration, e.g. maximum of the F_0. Fig. 4.14 shows examples of the F_0 features. In addition, 16 global prosodic features for the whole utterance are calculated (cf. Table 4.3). They cover each

Turn level features	
Feature	Description
F0MeanGlobalWord	Mean of the F_0 value in the current utterance
F0VarianceGlobalWord	Variance of the F_0 value in the current utterance
Mean_jitter	Mean value of the jitter in the current turn
Variance_jitter	Variance of the jitter in the current turn
Mean_shimmer	Average of the shimmer in the current turn
Variance_shimmer	Variance of the shimmer in the current utterance
Num_V_Segments	Number of voiced segments in the current utterance
Num_UV_Segments	Number of unvoiced segments in the current utterance
Len_V_Segments	Length of the voiced segments in the current turn
Len_UV_Segments	Length of the unvoiced segments in the current turn
MaxLen_V_Segments	Maximal length of a voiced segment in the current utterance
MaxLen_UV_Segments	Maximal length of an unvoiced segment in the current utterance
RatioNum_VUV_Segments	Ratio of the number of voiced and unvoiced segments in the current turn
RatioLen_VUV_Segments	Ratio of the length of voiced and unvoiced segments in the current turn
RatioLen_VSignal_Segments	Ratio of the length of the voiced segments and the current utterance
RatioLen_UVSignal_Segments	Ratio of the length of the unvoiced segments and the current utterance

Table 4.3: Overview on the prosodic features computed on turn level

of mean and standard deviation for jitter and shimmer, the number, length and maximum length each for voiced and unvoiced sections, the ratio of the numbers of voiced and unvoiced sections, the ratio of length of voiced sections to the length of the signal and the same for unvoiced sections. Jitter and shimmer are extracted as described in [Levit 00, p.14]. The last global feature is the variance of the fundamental frequency F_0. In order to evaluate pathologic speech, we calculate the average, the maximum, the minimum, and the variance of the 37 turn- and word-based features for the whole text to be read. Thus, we get 148 features for the whole text.

The mean value F0MeanGlobalWord is computed for a window of 15 words (or less if the utterance is shorter) [Batli 99, Batli 01] so it is regarded as turn level feature here.

In contrary to features of many other research groups, our features do not make a hard decision: instead of ‘stylizing’ the F_0 contour (‘hat contour’, ‘rise’, ‘rise fall’,

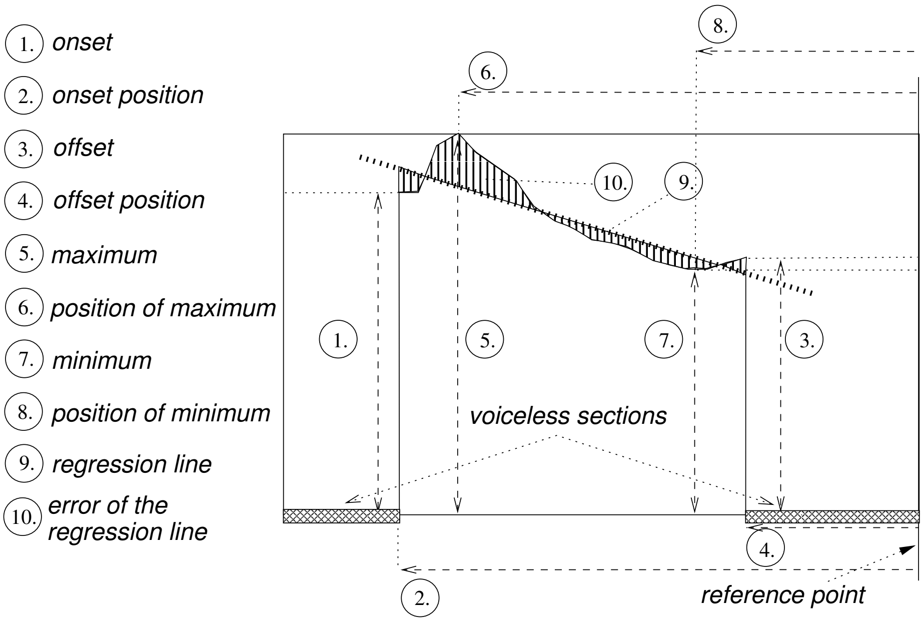


Figure 4.14: Computation of prosodic features within one word (after [Kiess 97])

'high tone', ...), we extract features such as Min, MinPos, Max, and MaxPos which implicitly describe the F_0 and also the energy contour and leave the decision to the classifier.

The features proved to be effective for linguistic and emotion analysis [Batli 03a, Huber 02], the detection of boundaries between phrases [Batli 95], the user state, such as "tired" or "not tired" [Adelh 03, Batli 03b], and the focus of attention [Hacke 06].

Pronunciation Analysis

For the analysis of the pronunciation, further features are extracted from the speech signal. Unlike the prosodic features, the pronunciation features are designed to model variations in speech which are typically caused by altered pronunciation of non-native speakers. This should also be applicable to pathologic speech. Different approaches are pursued in this work: Features which are independent of the actual language and language-dependent features.

To model pronunciation variations language-independently, some kind of prior information has to be implemented to the feature extraction process. This is performed either data-driven or by the use of heuristics in order to model specific characteristics of speech:

- **PronFex:** Pronunciation features, as described in [Hacke 05b], were designed to rate a speaker's pronunciation in a data-driven approach. They are used for measuring the progress in learning a foreign language [Cinca 04b]. In this work, we study these features' applicability to the detection of pathologic speech. More precisely, we only analyze a subset of these features that is based on phoneme confusion probabilities on word level. To calculate these phoneme confusion features, we compare the result of the forced alignment with the Viterbi algorithm (cf. Chapter 4.2.4) of every word to the result of a phoneme recognizer. The phoneme recognizer uses semi-continuous hidden Markov models as described in Chapter 4.2.4 and a 4-gram language model (cf. Chapter 4.2.4). It is based on MFCCs calculated every 10 ms with a frame size of 16 ms. From these informations phoneme confusion matrices $\boldsymbol{C}$ are built. They contain for every pair of phonemes a, b the probability that a was detected by the recognizer when there should be b according to the forced alignment

$$c_{ab} = P(a \mid b) \tag{4.122}$$

where c_{ab} is the corresponding entry of matrix $\boldsymbol{C}$. From the training set, we calculate two confusion matrices: one for the pathologic speech data and one for the normal data. The quotient Q is calculated for every frame:

$$Q = \frac{P_{\text{pathologic}}(a \mid b)}{P_{\text{normal}}(a \mid b)} \tag{4.123}$$

From these frame-wise results, we calculate the following features for the phone level [Cinca 04a, p.162]:

- Goodness of Pronunciation (GOP): Likelihood obtained by a GMM classifier (cf. Chapter 4.2.4) which was trained with speech of the target language. In non-native speech, the likelihood is known to drop in mispronounced phones.
- Duration Score: Probability of the observed phone duration, given the duration distribution observed in native speakers
- Acoustic Score: Confidence of the speech recognizer for the current phone (cf. Chapter 4.2.4)
- Confidence Score: Q
- Actual Duration: Observed duration
- Expected Duration: Mean value of the duration distribution observed in native speakers

For word level the following features are extracted [Cinca 04a, p.162]:

- PC1: Mean of Q
- PC2: Maximum of Q
- PC3: Minimum of Q
- PC4: Variance of Q
- PC5: Median of Q
- A1: Phoneme correctness
- A2: Confidence score of the recognized word, computed by the speech recognizer (cf. Chapter 4.2.4)

- **Nasality Detection using the Teager Energy Operator:** The Teager Energy Operator is a heuristic approach of language-independent pronunciation feature extraction. The Teager Operator [Teage 90] has been applied to detect nasality in sustained vowels and consonant-vowel-consonant combinations [Cairn 96b]. The Teager Energy operator (TEO) is defined as:

  $$\psi[f(n)] = [f(n)]^2 - f(n+1)f(n-1) \tag{4.124}$$

  $f(n)$ denotes the time-domain audio signal. The TEO's output is called the Teager Energy Profile (TEP).

  As described in [Cairn 96a] and [Cairn 96b], the TEP can be used to detect hypernasal speech because it is sensitive to multicomponent signals. When normal speech is low-pass-filtered in a way that the maximum frequency f_{lowpass} is somewhere between the first and the second formant, the resulting signal mainly consists of the first formant. However, doing the same with hypernasal speech results in a multicomponent signal due to the anti-formant (cf. Eq. 3.3). If we now compare the low-pass-filtered TEP to the TEP of the same signal that was bandpass-filtered around the first formant, we should see more difference in case of a hypernasal signal. We measure that difference with the correlation coefficient (cf. Eq. 4.1) of the TEPs. The bandpass filter covers the frequency range ±100 Hz around the first formant. The values with the best results for f_{lowpass} were determined experimentally and are listed in [Reuss 07].

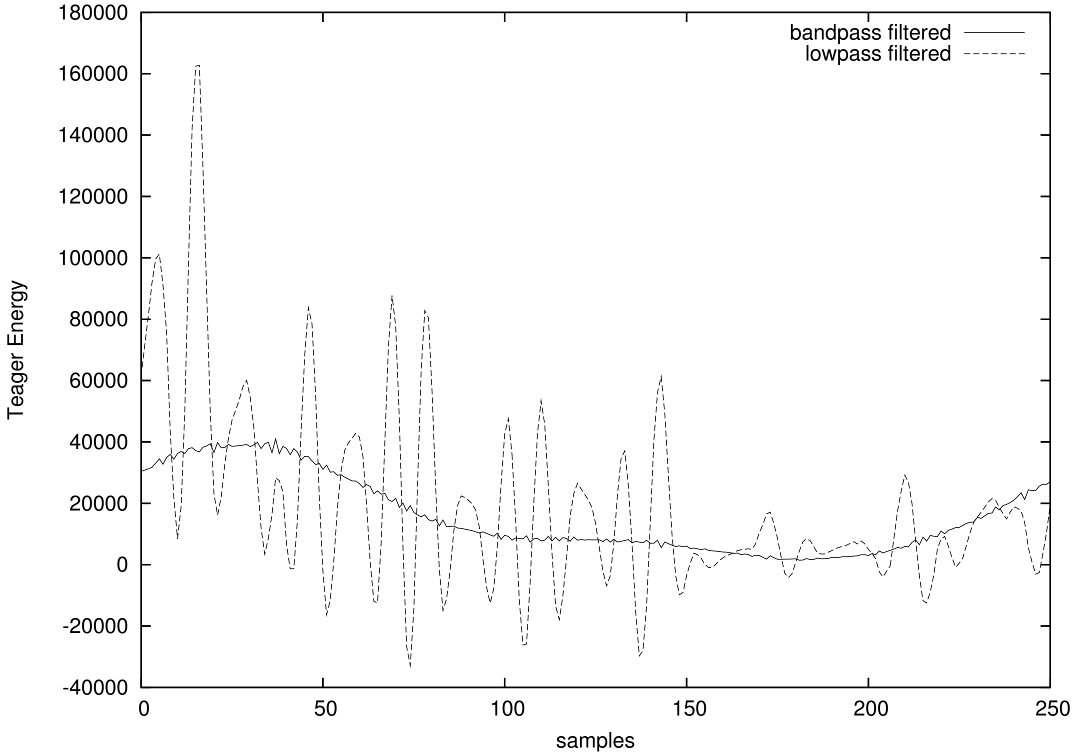
  Figure 4.15 shows examples for this. Figure 4.15(a) displays a low-pass filtered and a bandpass-filtered TEP of the same audio frame with a rather high correlation, and Figure 4.15(b) shows the same for a different audio frame with a low correlation between both TEPs. Figure 4.15(a) represents what we expect to observe in case of normal speech while Figure 4.15(b) demonstrates the hypernasal case. Both figures show frames classified as an /i:/ (SAMPA notation, cf. [Wells 97]).

For the language-dependent case, an interesting approach is stated in [Hacke 07b]. The idea is to model the pronunciation variants which occur often as additional entries in the lexicon of a speech recognizer. The speech recognizer selects the best-fitting model form the lexicon during the decoding process (cf. Chapter 4.2.4). This yields an integrated recognition of pronunciation errors during the speech recognition process.

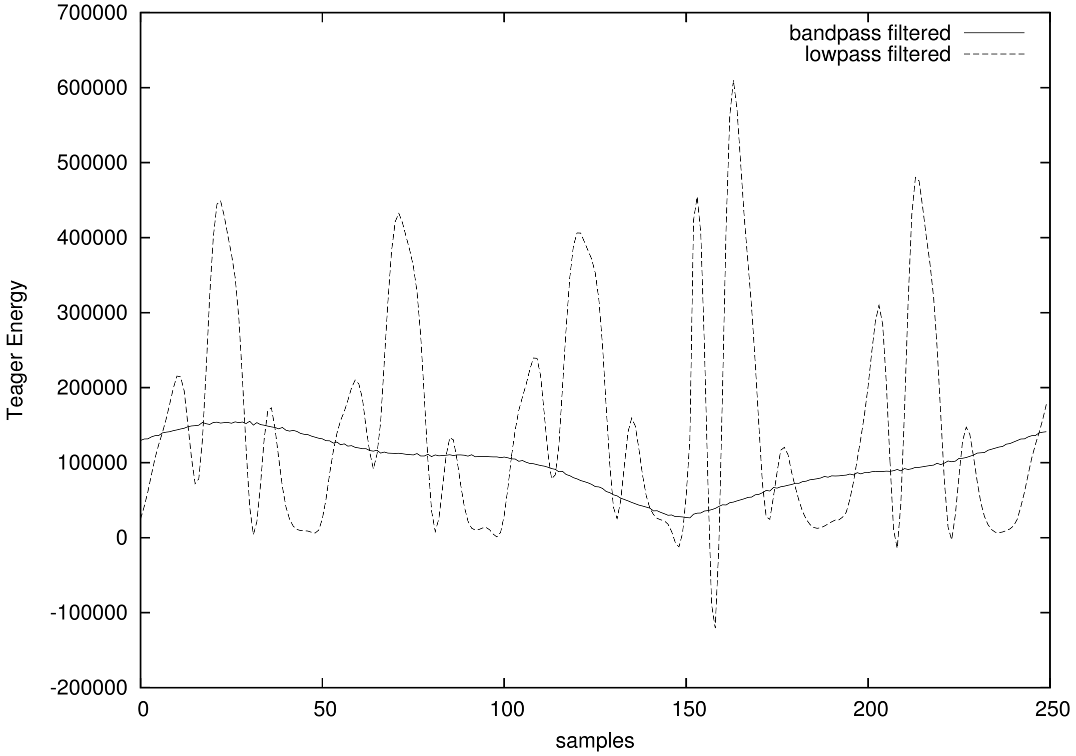
However, this method has a major drawback: All pronunciation variants have to be determined before the recognition process. In [Hessl 05], for example, this has been done in cooperation with a teacher for the English language. From the manually created pronunciation variant list, a set of rules was derived which was then applied to a vocabulary list in order to create further pronunciation variants.

4.2.5 Normalization of Age Effects

For the recognition of children's speech, the age plays an important role. In general, the speech of children is more variable than adults' speech [Wilpo 96]. In order to cope with this variance, the speaker-adaptation of the speech recognizer is beneficial [Maier 06b]. Furthermore, the anatomy of children differs from adults. The children's



(a) high correlation (normal case)



(b) low correlation (hypernasal case)

Figure 4.15: Examples of high and low correlation between a low-pass- (1900 kHz) filtered and a bandpass-filtered TEP

vocal tract and the vocal cords are shorter than in adults. This results in higher fundamental and formant frequencies in the speech of children. Using *Vocal Tract Length Normalization* (VTLN, cf. [Eide 96]), this effect can be attenuated.

Speaker Adaptation

The Gaussian densities of the speech recognizer can be adapted similar to the MAP adaptation described in Chapter 4.2.4. In addition, improvement can be achieved by *Maximum Likelihood Linear Regression* (MLLR) adaptation. The idea is to find a set of linear transformation matrices $\boldsymbol{W}$ for the mean values and $\boldsymbol{H}$ for the covariance matrices based on the ML estimate. In order to perform the adaptation, these matrices are applied in the following manner:

$$\begin{aligned} \hat{\boldsymbol{\mu}_m} &= \boldsymbol{A}\boldsymbol{\mu}_m + \boldsymbol{b} = [\boldsymbol{A}\boldsymbol{b}] \begin{bmatrix} \boldsymbol{\mu}_m \\ 1 \end{bmatrix} = \boldsymbol{W}\boldsymbol{\xi}_m && (4.125) \\ \hat{\boldsymbol{\Sigma}}_m &= \boldsymbol{H}\boldsymbol{\Sigma}_m\boldsymbol{H}^\top && (4.126) \end{aligned}$$

$\boldsymbol{\xi}_m$ denotes the extended mean vector which is found by concatenating 1 to the mean vector $\boldsymbol{\mu}_m$. $\boldsymbol{W}$ consists of a transformation matrix $\boldsymbol{A}$ and a translation vector $\boldsymbol{b}$. The covariance $\boldsymbol{\Sigma}_m$ is adapted by applying $\boldsymbol{H}$. According to [Gales 97] the matrices can be computed by a version of the EM algorithm. In the expectation step, a likelihood function $\mathcal{L}(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}})$ is computed. In the most general form, this can be written as

$$\mathcal{L}(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}}) = K_0 - \frac{1}{2}\sum_m\sum_t \gamma_t(m)\left[K_m + \log\left(\left|\hat{\boldsymbol{\Sigma}}_m\right|\right) + (\boldsymbol{x}_t - \hat{\boldsymbol{\mu}}_m)^\top \hat{\boldsymbol{\Sigma}}_m^{-1} (\boldsymbol{x}_t - \hat{\boldsymbol{\mu}}_m)\right] . \tag{4.127}$$

$\boldsymbol{\lambda}$ denotes the original models while $\hat{\boldsymbol{\lambda}}$ are the adapted models. $\hat{\boldsymbol{\Sigma}}_m$ and $\hat{\boldsymbol{\mu}}_m$ are the adapted mean and covariance. K_0 is a constant only dependent on the transition probabilities, and K_m is the normalization constant corresponding with the Gaussian component m. The $\gamma_t(m)$ give the probability to select Gaussian component m at time t which can be computed with the Baum-Welch algorithm from Chapter 4.2.4. In Eq. 4.109 only $\gamma_t(j)$ is computed which can be used to compute $\gamma_t(m)$ by

$$\begin{aligned} P(q_t = s_j, k_t = m|\boldsymbol{x}^T, \boldsymbol{\lambda}) = &\ \gamma_t(j)\frac{c_{jm}\mathcal{N}_m(\boldsymbol{x}_t)}{\sum\limits_l^M c_{jl}\mathcal{N}_l(\boldsymbol{x}_t)} = \zeta_t(j, m) && (4.128) \\ \gamma_t(m) = &\ \sum_j \zeta_t(j, m) \ . && (4.129) \end{aligned}$$

In the maximization step of the algorithm, the likelihood function is maximized. The problem is simplified by several restrictions to the transformation of $\hat{\boldsymbol{\mu}}_m$ and $\hat{\boldsymbol{\Sigma}}_m$. In the following mean transformation and covariance transformation are presented.

- **Mean Vector Transformation:** In order to transform the mean vector, the transformation matrix $\boldsymbol{W}$ is applied in the following manner:

$$\hat{\boldsymbol{\mu}}_m = \boldsymbol{A}\boldsymbol{\mu}_m + \boldsymbol{b} = \boldsymbol{W}\boldsymbol{\xi}_m \tag{4.130}$$

The estimate of the covariance matrix $\hat{\boldsymbol{\Sigma}}_m$ is set to $\boldsymbol{\Sigma}$. Thus, the likelihood function can be simplified to

$$\begin{aligned}\mathcal{L}(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}}) = \;& K_0 - \frac{1}{2} \sum_m \sum_t \gamma_t(m) \Big[K_m + \log\left(|\boldsymbol{\Sigma}_m|\right) \\ & + \left(\boldsymbol{x}_t - \hat{\boldsymbol{\mu}}_m\right)^\top \boldsymbol{\Sigma}_m^{-1} \left(\boldsymbol{x}_t - \hat{\boldsymbol{\mu}}_m\right) \Big] \,.\end{aligned} \tag{4.131}$$

The following system of equations is obtained by setting the derivative to zero in the M step:

$$\sum_t \sum_m \gamma_t(m) \boldsymbol{\Sigma}_m^{-1} \boldsymbol{x}_t \boldsymbol{\xi}_t^\top = \sum_t \sum_m \gamma_t(m) \boldsymbol{\Sigma}_m^{-1} \boldsymbol{W} \boldsymbol{\xi}_t \boldsymbol{\xi}_t^\top \tag{4.132}$$

Unfortunately, $\boldsymbol{W}$ is inside a sum over t and m which is computationally expensive. In [Gales 96] the problem is solved by calculation in variance-normalized domain. The inverse of the covariance matrix $\boldsymbol{\Sigma}_k^{-1}$ is decomposed to its *Cholesky* factor $\boldsymbol{C}_m$:

$$\boldsymbol{\Sigma}_m^{-1} = \boldsymbol{C}_m \boldsymbol{C}_m^\top \tag{4.133}$$

Cholesky factorization can be done on any symmetric positive definite matrix [Wilki 71]. The whole system can be normalized by application of the Cholesky factor which enables a fast computation of $\hat{\boldsymbol{\mu}}_m$:

$$\tilde{\boldsymbol{\mu}}_m = \boldsymbol{C}_m^\top \boldsymbol{\mu}_m \tag{4.134}$$

$$\hat{\boldsymbol{\mu}}_m = \left[\boldsymbol{C}_m^\top\right]^{-1} \tilde{\boldsymbol{A}} \boldsymbol{C}_m^\top \boldsymbol{\mu}_m + \boldsymbol{b} \tag{4.135}$$

$$\breve{\boldsymbol{\mu}}_m = \tilde{\boldsymbol{A}} \tilde{\boldsymbol{\mu}}_k + \tilde{\boldsymbol{b}} = \left[\tilde{\boldsymbol{A}} \tilde{\boldsymbol{b}}\right] \begin{bmatrix} \tilde{\boldsymbol{\mu}}_m \\ 1 \end{bmatrix} = \tilde{\boldsymbol{W}} \tilde{\boldsymbol{\xi}}_m \tag{4.136}$$

The matrices $\tilde{\boldsymbol{W}}$ and $\tilde{\boldsymbol{A}}$ and the vectors $\tilde{\boldsymbol{b}}$, $\tilde{\boldsymbol{\mu}}_m$, and $\tilde{\boldsymbol{\xi}}_m$ denote the respective matrices and vectors in variance-normalized domain. Now the maximization in normalized domain for $\breve{\boldsymbol{\mu}}_m$ yields a system of equations similar to Eq. 4.132:

$$\begin{aligned}\sum_t \sum_m \gamma_t(m) \boldsymbol{C}_m^\top \boldsymbol{x}_t \tilde{\boldsymbol{\xi}}_m^\top &= \sum_t \sum_m \gamma_t(m) \tilde{\boldsymbol{W}} \tilde{\boldsymbol{\xi}}_k \tilde{\boldsymbol{\xi}}_m^\top \\ &= \tilde{\boldsymbol{W}} \sum_t \sum_m \gamma_t(m) \tilde{\boldsymbol{\xi}}_m \tilde{\boldsymbol{\xi}}_m^\top\end{aligned} \tag{4.137}$$

However, this system can be resolved easily in order to compute $\tilde{\boldsymbol{W}}$. Then Eq. 4.135 can be applied to find the transformed MLLR estimate $\hat{\boldsymbol{\mu}}_k$ outside the normalized domain.

- **Covariance Transformation:** According to [Gales 96], for MLLR covariance adaptation the covariance $\boldsymbol{\Sigma}_m$ is transformed using the transformation matrix $\boldsymbol{H}$ as defined in Equation 4.126. Again, the Cholesky factor $\boldsymbol{C}_m$ of $\boldsymbol{\Sigma}_m^{-1}$ from Eq. 4.133 is applied to transform the system of equations into the normalized domain:

$$\hat{\boldsymbol{\Sigma}}_m = \left[\boldsymbol{C}_m^\top\right]^{-1} \hat{\boldsymbol{H}} \boldsymbol{C}_m^{-1} \tag{4.138}$$

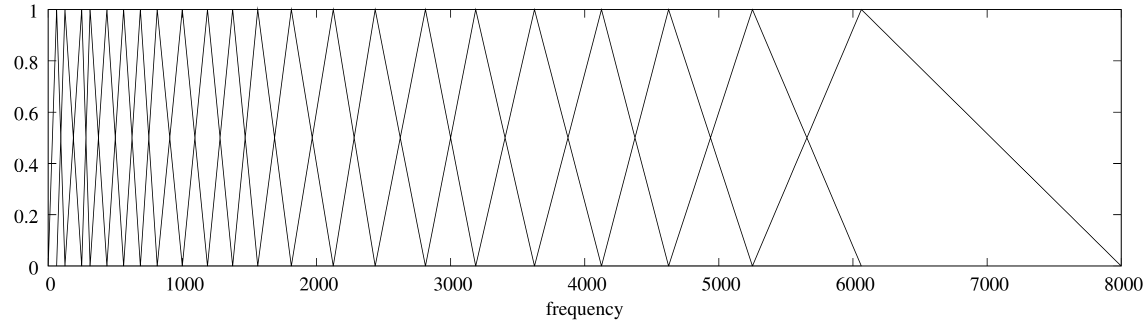


Figure 4.16: The composition of the Mel filter bank with 22 filters after VTLN with a scaling factor of $\frac{1}{\nu} = 1.2$ (cf. Figure 4.11 for the configuration before the VTLN)

Using this definition the following likelihood function $\mathcal{L}(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}})$ is derived from Eq. 4.127:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}}) &= K_0 - \frac{1}{2}\sum_m \sum_t \gamma_t(m)\Big[K_m + \log\left(\left|\left[\boldsymbol{C}_m^\top\right]^{-1} \hat{\boldsymbol{H}}\,\boldsymbol{C}_m^{-1}\right|\right) \qquad (4.139)\\ &\quad + (\boldsymbol{x}_t - \hat{\boldsymbol{\mu}}_m)^\top \left[\boldsymbol{C}_m^\top\right]^{-1} \hat{\boldsymbol{H}}\,\boldsymbol{C}_m^{-1}\,(\boldsymbol{x}_t - \hat{\boldsymbol{\mu}}_m)\Big]\end{aligned}$$

This estimate is maximized by computation of the derivative and setting it to zero. The solution of the emerging system of equations is found as

$$\hat{\boldsymbol{H}} = \frac{\sum_m \left\{\boldsymbol{C}_m^\top \left[\sum_t \gamma_t(m)\,(\boldsymbol{x}_t - \hat{\boldsymbol{\mu}}_m)\,(\boldsymbol{x}_t - \hat{\boldsymbol{\mu}}_m)^\top\right] \boldsymbol{C}_m\right\}}{\sum_m \sum_t \gamma_t(m)}. \qquad (4.140)$$

Note that the estimate $\hat{\boldsymbol{\mu}}_m$ of the mean vector $\boldsymbol{\mu}_m$ is necessary to compute $\hat{\boldsymbol{H}}$. Thus, the $\hat{\boldsymbol{\mu}}_m$ have to be computed in a first pass on the adaptation set according to Eq. 4.135. In a second pass over the whole adaptation set, $\hat{\boldsymbol{H}}$ is found afterwards.

Vocal Tract Length Normalization

VTLN is used to warp the length of the vocal tract of one person to the length of another person. In our case we desire to warp either children's speech to adults' speech or vice versa.

In [Eide 96] a linear approach for VTLN is presented. According to [Fant 73] the frequency of the formants are dependent on the length of the vocal tract L. The formant frequency F_k of the k^{th} formant is directly proportionate to $\frac{k}{L}$. Thus, a linear scaling of the vocal tract length according to a scaling factor ν yields $L' := L\nu$:

$$F_k' \propto \frac{k}{L'} = \frac{k}{L\nu} = \frac{1}{\nu}\frac{k}{N} \propto \frac{1}{\nu}F_k \qquad (4.141)$$

Hence, a linear scaling of the vocal tract length L corresponds to a linear scaling of the frequency axis with $\frac{1}{\nu}$. Such a linear VTLN can easily be integrated into the feature extraction process described in Chapter 4.1.3 by adaptation of the filter banks. Figure 4.16 shows the configuration of the filter banks after a scaling of $\frac{1}{\nu} = 1.2$

(cf. Figure 4.11 for the configuration before the VTLN). In [Stemm 05] an overview on further VTLN methods is given.

4.3 PEAKS Architecture

In this section the architecture of PEAKS is described in detail. As already shown in Figure 4.2, PEAKS can be divided into three major blocks: The client, the transport layer, and the server.

4.3.1 Classes of the PEAKS Client

Most of the classes in the client block are part of the graphical user interface. The corresponding methods are embedded in the respective part of the graphical user interface. Since PEAKS employs Swing[4] for the graphical user interface, this part also uses the Swing terminology, i.e. “JApplet” refers to a Java^{TM} program which is executed as a part of a web page, “JFrame” is a window which is opened additionally by the applet, and “JPanel” is a graphical user interface which can either be embedded into a JApplet, a JFrame, or a JPanel.

Tables 4.4, 4.5, and 4.6 give an overview on the most important classes in the PEAKS client. The tables are grouped according to their functionality: The administration tools (cf. Table 4.4) can only be opened by a user who is flagged as administrator in the system. Furthermore, a special administrator version of PEAKS which is not available on the Internet is necessary to gain access to these functions. In the following the classes which handle the PEAKS client functionality (cf. Table 4.5) are described. PEAKS provides classes to register new users and patients. Furthermore, patients can be recorded using various different tests depending on the type of patients and the intended type of examination. Additional applets (cf. Table 4.6) provide the functionality to create recordings easily from the home of the patient without having to use the full PEAKS client software. Another additional applet can be employed to test whether the client PC fulfills all technical requirements to use PEAKS. A detailed description of all functionalities of the PEAKS client is given in the PEAKS manual [Maier 06a].

[4] For a documentation of Swing, please go to http://java.sun.com/docs/books/tutorial/uiswing/

Administration tools		
super class	name	description
JFrame	AdminCommandFrame	This frame gives an administrator of the system access to various functions such as management of access control for the users or the modification of PEAKS internal pronunciation lexicon.
JFrame	GroupAccessControlFrame	A frame to control the access permissions of the individual users.
JFrame	LexiconEditor	A frame to manipulate the pronunciation lexicon of PEAKS.

Table 4.4: Administration classes of the PEAKS client

PEAKS client functionality		
super class	name	description
JApplet	Peaks	The main applet of PEAKS all panels are linked to this applet in order to be displayed. By default, the Peaks applet shows the login screen.
JPanel	UserRegisterPanel	The panel which is used to register new users to the system.
JPanel	MainMenu	The main menu GUI which gives access to all functionalities of PEAKS
JFrame	LogFrame	A frame which logs all important messages of PEAKS. It is opened if the web browser does not provide access to the Java$^{\mathrm{TM}}$ console.
JPanel	PatientRegisterPanel	This panel is used to add new patients to the PEAKS system.
JPanel	PatientEditPanel	A panel to edit the information on a patient in the system.
JPanel	PatientLabelPanel	This panel is used to enter subjective evaluations into the PEAKS system.
JPanel	RecordSelection	In this panel the test to be recorded is chosen. Furthermore, it enables the user to create a “RecordLink” which can be sent via e-mail to a patient to perform the recording from his home.
Runnable	APstarter	A process which can be applied to create batch recordings of multiple tests right after each other.
JPanel	AudioRecPanel	The actual recording panel. It can display a sequence of pictograms or texts or both at the same time.
JPanel	ExpertLabelingPanel	The panel which is used to enter the detailed assessment of a speech therapist.
Runnable	FileDataTransmission	A process which buffers the audio data for the transmission. In this manner PEAKS can also be used from computers with a slow Internet connection.
JPanel	TransmissionHandler	This panel displays the progress of the transmission after the recording was performed.
JPanel	RecordInfoPanel	A panel to supply further information about a recorded test.
JPanel	TranscriptionPanel	Using this panel the audio data can be transcribed.
abstract Object	Utils	An abstract class which contains common methods which are used in PEAKS.
abstract Object	ClientTransfer	Class to handle the communication with the server.
Interface	ReturnApp	An interface which enables to send information between different panels.
Interface	ReturnAppUpdateName	An extension of the ReturnApp interface to select a specific user in the main menu.

Table 4.5: Main classes of the PEAKS client

Additional PEAKS applets		
super class	name	description
JApplet	SimpleExerciseRecorder	The applet which is used to perform the "RecordLink" recording.
JApplet	PeaksTest	A small applet which checks for the correct Java^{TM} version and port blocks in the firewall of the client system.

Table 4.6: Additional classes of the PEAKS client

4.3.2 Classes of the PEAKS Transport Layer

A crucial part in the PEAKS system are the classes of the transport layer. In the following we describe the transfer objects which are employed in the transfer and the classes that handle the transfer.

While basically all classes from the previous section were part of the application layer, the classes of the transport layer can be divided into transfer objects which are used by the client and the server side to exchange information and transfer handlers. On the client side, all transfers are managed by the class "TransferHandler"; on the server side the corresponding class is "ServerThread". The transfer objects are listed in Table 4.7. All transfer objects implement the Java^{TM} interface "Serializable" which enables the objects to be written into a stream. Result objects also implement the interface "SQLData" since they have to be written into an SQL database.

A special role in the data exchange between the client and the server play the User and Session objects since they are used in order to set up the connection. In order to log into the system, the client sends first a User object, where only the member variables "name" and "password" (MD5-encrypted [Rives 92]) are set. All other member variables are set to null. The server checks now whether the password matches the password in the database and returns a Session object if the password is correct or aborts the connection if the password did not match. All subsequent connections are now initialized with the Session object instead of the username and the password since the Session object can be used to identify the user. Furthermore, the Session object has an expiry time which is only stored on the server side in order to avoid manipulation by the client. Session objects which do not come from the same host as the initial password request came from are discarded.

The transfer objects are used for all communication in PEAKS. The objects can be used as both result and request. All data transfers in PEAKS follow the general convention that "empty" objects, i.e., objects with null reference variables, are regarded as requests while "full" objects are regarded as results. If the client, for example, desires more information about a specific patient but knows just the ID of the patient, the client will produce a new User object with the corresponding patient ID. The server will then check whether the requesting user has access to the patient data and will return a User object with all member variables filled if the check was successful. All other communication in PEAKS is performed in the same manner.

Transfer objects		
super class	name	description
Object	AccessControl	Object to store access control information.
Object	Command	Object to implement special commands into the PEAKS transfer layer.
Object	Exercise	Object to represent a test to be recorded. An exercise consists of multiple turns.
Object	Turn	One sentence or word within an exercise.
Object	FileData	Representation of the audio data which was recorded in one turn.
Object	PatientContext	A context group of patients, e.g. children with cleft lip and palate.
Object	Result	Object to represent an arbitrary result obtained for one or multiple FileData objects.
Object	SammonMap	Representation of a visualization of a patient group.
Object	Session	Object to represent a PEAKS session.
Object	User	Object to represent users and patients of the PEAKS system.
Result	IntelligibilityResult	Representation of the result of the subjective evaluation of one FileData object by one user.
Result	ExpertRating	Result of the detailed assessment of one FileData object by an expert listener.
Result	TranscriptionResult	Transcription of one FileData object performed by a User.
Result	Lexikon	Object to store the global pronunciation lexicon.
Result	DoubleValueResult	Object to store an arbitrary double value as result.
Result	ProsodicFeatures	Result type to store the outcome of the prosodic analysis for a whole test.
Result	WAResult	Result type to store the recognition result of a test.
Result	WordHypothesisGraph	Result to store the time alignment performed by the speech recognizer for one FileData object.
Result	WordPhoneAlignment	Result to store the phone time alignment within one word. This object is used in WordHypothesisGraph to store the time alignment information of the individual words.

Table 4.7: Overview on the transfer objects in the PEAKS transfer layer

PEAKS main server classes		
super class	name	description
Object	PeaksServer	The main class of the server. It is used to start the server.
Runnable	ServerThread	The thread which processes a request of a PEAKS client
Runnable	ServerThreads	The thread which accepts new connections and spawns new ServerThread objects.
abstract Object	Definitions	Class to provide constant values like frame length and sampling rate to the PEAKS server.

PEAKS auxiliary server classes		
super class	name	description
Object	Database	Gateway to the database. Provides all functions to access the database.
Object	DocumentWrapper	Class to generate PDF result sheets.
Object	RecognitionWrapper	Class to provide access to the underlying code of the speech recognizer and other assessment methods as described in Chapter 4.2.
Object	ImportVoiceTest	Class to import new tests to PEAKS from text files.

Table 4.8: Overview on the classes in the PEAKS server

4.3.3 Classes of the PEAKS Server

The PEAKS server consists of four main server classes and four auxiliary classes (cf. Table 4.8). The main class “PeaksServer” is used to start the PEAKS server. It will spawn a “ServerThreads” object which starts listening on port 7070. If a client connects to that port, a new “ServerThread” is spawned to handle the request of the client. After the request is processed, the “ServerThread” is removed from the memory by the Java^{TM} garbage collection. An abstract class “Definitions” holds the constant variables, such as the frame length or the sampling rate of the audio data.

The auxiliary classes are used to gain access to certain information sources or sinks. The class “Database” provides access to a MySQL database which should be running in the same network as the PEAKS server. It provides all functions to store and load all types of transfer objects into the database. While all transfer objects which are of the type “Object” are stored in individual tables, the “Result” types are stored in the table “result” as binary large objects (BLOBs)[5]. This procedure is very convenient to implement, but it also has the disadvantage that any changes to the structure of the result object causes a modification in the serialization of the object. Hence, special serialization procedures have to be written manually, or the changes to the result objects have to be performed in inherited classes.

[5] In fact, the type MEDIUMBLOB is used since BLOB defaults to 65535 characters which is not as large as the term “binary large object” suggests.

The class “DocumentWrapper” is used to create PDF documents such as reports about the patients. Therefore, a LaTeX document is created and compiled. The report includes visualization, recognition performance, and other assessment results in comparison to a patient group. These reports can then be sent to the client where they can be reviewed.

In order to compute the various assessment and evaluation procedures which are offered by PEAKS, the class “RecognitionWrapper” is employed. In the PEAKS server, there is just a single instance of the “RecognitionWrapper”. All tasks to be performed by it are stored in a queue and are processed in a first-in-first-out (FIFO) manner since some parts of the speech recognition engine cannot run concurrently. Furthermore, there is no need for real-time processing of the data since all processing is offline. As long as there are still tasks in the queue of the wrapper, it will work until all tasks are finished. The respective results are stored in the database where they can easily be accessed by the client. With the “RecognitionWrapper” it is possible to run all the computationally intensive analyses described in Chapter 4.2 in C or C++ while the client written in Java^{TM} is platform-independent.

The class “ImportVoiceTest” offers methods to integrate new voice and speech tests into PEAKS easily. Therefore, a test file in a specific format has to be uploaded to a publicly accessible web server. If the test should contain pictures as well, these must also be uploaded to the web server. Then the URL of the file can be supplied to the import routine. This causes the import routine to create all corresponding entries in the database, thus enabling recordings of the test. In order to enable automatic evaluation, the pronunciation of all words of the test has also to be supplied to the recognition system (cf. Chapter 6).

Chapter 5

Data Collection

This chapter reports all data which were collected during the work of this thesis. All data were collected during a German speech test – the Psycho-Linguistic Analysis of Children's Speech Disorders. Reference data were collected in several areas of Germany. About 800 control children were collected throughout this work. Furthermore, about 400 children with cleft lip and palate were recorded at the University Hospital in Erlangen. Informed consent had been obtained by all parents prior to the recordings.

5.1 Psycho-Linguistic Analysis of Children's Speech Disorders

The Psycho-Linguistic Analysis of Children's Speech Disorders is a semi-standardized test commonly used by speech therapists. The test is called "Psycho-Linguistische Analyse kindlicher Sprechstörungen" in German and abbreviated as PLAKSS [Fox 02]. It is designed for the assessment of speech disorders in children aging between 4 and 18 years. Some of the children who are tested are not yet able to read. Hence, the test consists of pictograms. During the test the speech therapist shows the pictograms to the child and encourages it to say the names of the presented items. The test consists of 99 words on 33 slides. 97 of the words are disjoint. Two words appear twice ("Ball" and "Pilz"). The vocabulary of the PLAKSS test can be reviewed in Appendix A.1.1. All German phonemes are included in the test. The German phonemes are tested in beginning, center, and end position of a word. Vowels, however, are not targeted in the test.

Figure 5.1 shows an example of the slides. Slide 13 consists of the German words "Trecker, Zitrone, Jäger" which mean tractor, lemon, and hunter in English. The slide gives a good example of the test: While the tractor and the lemon are quite easy to identify, the hunter often poses a problem. Many children do not recognize the rifle on the back of the hunter and call the pictogram "man with a dog". Furthermore, the word "Trecker" is rather uncommon in the southern part of Germany. Children tend to prefer variants such as "Traktor" or "Bulldog". Therefore, the vocabulary of the PLAKSS test has to be extended with common word alternatives and regional variants, if their detection is desired. A list of all common word alternatives is

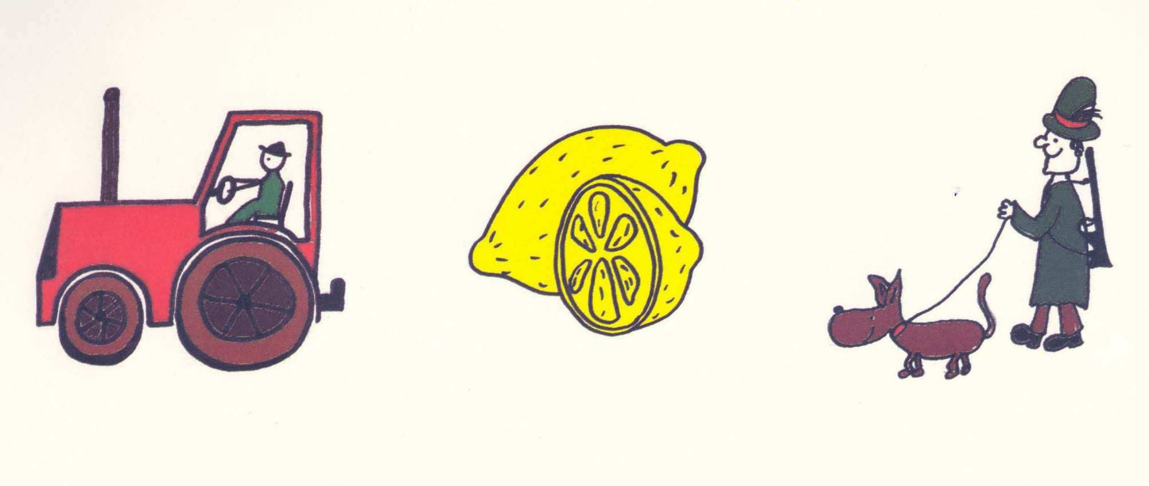


Figure 5.1: Slide 13 of the PLAKSS test: "Trecker, Zitrone, Jäger" (tractor, lemon, hunter)

presented in Appendix A.1.2. The assessment procedure, however, is only defined on the target phonemes given by the test. The forms for the perceptive assessment of the PLAKSS test are presented in Appendix A.2.

In the first version of the speech data collection software, the data were recorded by PC directly to a local harddisk. All the data of one test were stored in a single file. This procedure had several disadvantages. First of all, the data had to be transliterated completely before they could be further processed since no additional information is known. Another disadvantage was that no information about the speech therapist is known. Data collection, however, was performed by more than one speech therapist. Thus, speaker identification techniques could not be applied. The speech of the therapist had to be separated manually from the children's speech. All data of the first version were automatically segmented at long pauses into turns before the manual processing (cf. [Stemm 05]). Then the turns were listened to in order to exclude the speech therapist. The last step in the processing was the transliteration. Due to the high effort which is necessary during the segmentation and transliteration, the processing time was about ten times the time consumption of the actual recording in this first assessment.

In order to enable automatic evaluation of the test, the PEAKS software environment was created (cf. Chapter 4). A large advantage of PEAKS is that the slides of the pictograms are displayed with the same software as the data recording is performed. Therefore, the data can be segmented according to the slides. This procedure provides additional information which can be used in the evaluation process in order to provide fully automatic assessment. Furthermore, since the speech therapist has to log into PEAKS. Thus, he is also automatically identified. This information is then saved in order to simplify the automatic segmentation between the therapist and the child. Annotation of the perceptive evaluation is also part of the PEAKS system. Specific panels for the assessment using Likert-scales and phoneme level annotation are provided by the system. Screenshots of their graphical user interface are presented in Appendix A.3.

5.2 Control Groups

The collection of the control group speech data was performed in five different cities of Germany to test for dialectal influences. Non-native children's speech was excluded from the data. Therefore, the mother tongue of the child's mother and father were noted down. Later, the decision whether the child was native or non-native was decided perceptually. Most of the data were collected in Erlangen. In the following the datasets are described in detail. In order to provide a better overview on the data, the datasets are grouped by the city in which they were collected.

5.2.1 Erlangen (Franconia)

In Erlangen, data were recorded at two different locations: An elementary school — the Michael-Poeschke-Schule — and a kindergarten — the Erna-Zink-Kindergarten.

In the Michael-Poeschke-Schule, data collection was performed three years subsequently. For the time of the data collection, the school's library was dedicated as recording room. The room had dimensions of about $6\,m \times 5\,m \times 3\,m$. Two of the walls were covered by bookshelves. In the room no strong reverberation was audible. Furthermore, there were no noise sources such as active PC fans or air-conditioners. Since no Internet connection was available, PEAKSlocal (cf. Figure 4.3) was used to simulate the real PEAKS server. In the first year, in January 2006, data were recorded using a Sennheiser close-talking microphone (handgrip K3U with ME 80 head). Since the microphone is very sensible to background noises, they had to be reduced. The microphone was placed in a box with dimensions of about $1\,m \times 1\,m \times 1\,m$. The walls of the box were covered by foam rubber in order to absorb sounds. One side of the box had a hatch which could be opened in order to place microphone and laptop into the box. The box was placed on a table, and a chair was put in front of the hatch of the box. For the recording, the hatch was opened, and the child was sat on the chair in front of the box. The distance of the children's mouth to the microphone was approximately 20 *cm*. Since the sound card of the laptop produced noise around 50 Hz when the power adapter was connected — caused by the AC power supply — digital analog conversion was performed using the converter in a TASCAM DA-P1 DAT recorder. The data were then directly streamed onto the harddisk of the laptop using PEAKSlocal. Sampling was performed at 48 kHz during the recording but later on resampled at 16 kHz. Quantization was performed at 16 bit. Figure 5.2 shows this recording setup.

In order to keep the time of missed classes as short as possible, the children left the class one by one and were escorted to the recording room. After the test, each child was brought back to the classroom, and the next child could follow. In total, data of 89 children (46 female and 43 male) at an age of 8.8 ± 1.3 years were recorded with the PLAKSS test. The recordings were in good quality. Only slight background noises occurred. Most of them were caused by children playing in the courtyard of the school during the breaks between the lessons. These data are called *Michael-Poeschke-06* in the following.

One year later — in April 2007 — more data were gathered at the Michael-Poeschke-Schule. In order to simplify the recording setup, a different configuration

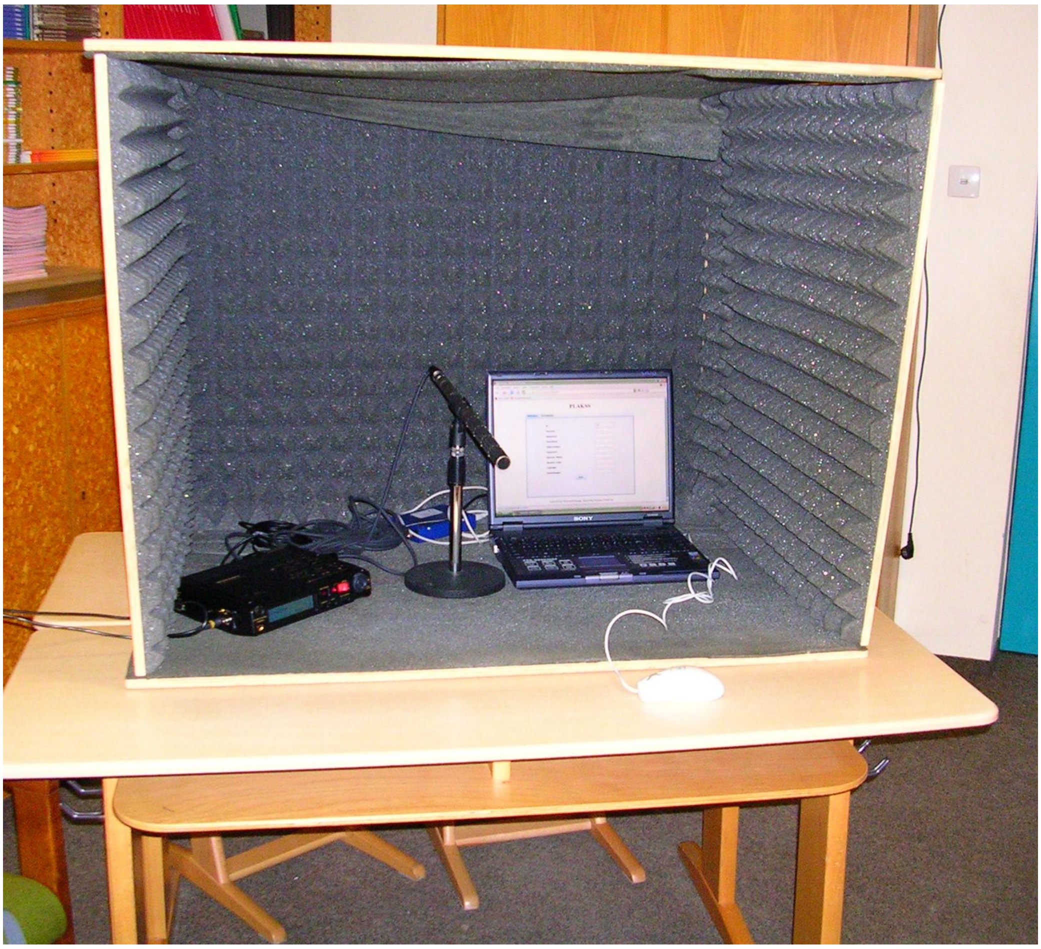


Figure 5.2: First setup of the recording environment

for the data collection was chosen. Instead of the recording box, the DAT recorder, and the Sennheiser microphone, we chose a USB head set (Plantronics Audio USB 510). So the recording laptop PC was placed on a table and a chair in front of it. Before the test each of the children had to put on the head set. Before each recording the microphone of the head set was adjusted to be in front of the child's mouth. The data were sampled at 16 kHz and quantized at 16 bit. Hence, no further resampling had to be performed. The recording quality of the head-mounted USB microphone was only slightly worse than the previous setup with the recording box. Due to the head-mounted microphone, fewer background noises were audible. Moreover, data collection was paused during the breaks between the lessons. Since the new recording configuration was much simpler and faster to deploy, and the audio quality was similar or only slightly worse, the setup was selected as the standard recording setup for the collection of control data. Figure 5.3 shows the simplified recording setup. Additional documentary images of the recording situation are presented in Appendix A.5. 76 children (39 female and 37 male) at the age of 8.5 ± 1.4 years were recorded. The data of the Michael-Poeschke-Schule collected in April 2007 are denoted *Michael-Poeschke-07* throughout this work.

In February 2008, a third time data were collected at the Michael-Poeschke-Schule. The recording setup, location, microphone, PC, and software were the same as in

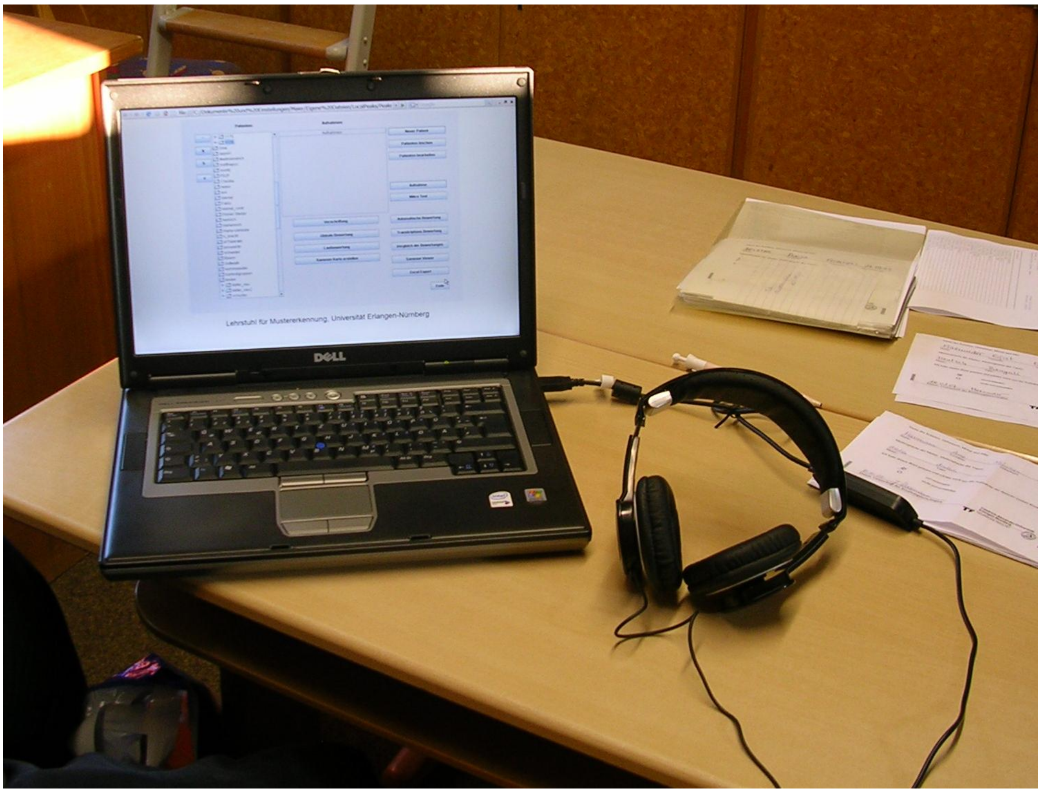


Figure 5.3: Simplified recording setup used since March 2007

2007. In the third year, 157 children (76 female and 81 male) at the age of 8.4 ± 1.2 years were recorded. The data are referred to as *Michael-Poeschke-08* in the following.

The youngest control group was recorded at a local kindergarten in Erlangen. 21 boys and 17 girls at the age of 5.7 ± 0.7 years were recorded with the PLAKSS in the Erna-Zink-Kindergarten. Additionally, the children were also looked at by a dentist in order to document missing teeth which might cause sigmatism. For the recording a USB head set (Plantronics Audio USB 510) was attached to a laptop PC (cf. Figure 5.3). No Internet connection was available. Therefore, PEAKSlocal was used again. As in all recordings performed with PEAKS, the sampling rate was 16 kHz and quantization 16 bit. The dataset is referred to as *Erna-Zink-07*.

5.2.2 Nuremberg (Franconia)

In Nuremberg, data were collected at a high school — the Sabel Schule. The school did also strongly support this work. They dedicated a room during the time of the recording. The room had the size of a normal classroom with about 50 m^2. The walls were even and plain. Reverberation was reduced by furniture, like tables and chairs for about 30 pupils. Using the new recording setup consisting of laptop PC and head set (cf. Figure 5.3), 48 children (20 female and 28 male) at the age of 13.2 ± 1.2 years were recorded in March 2007. The dataset is denoted as *Sabel-07* in the following.

5.2.3 Hannover (North Rhine-Westphalia)

In order to test whether our evaluation methods are independent of the dialect, data were collected all over Germany. Speech data representing standard German were collected in Hannover. The Eichendorff elementary school agreed to allow recordings at their school. They dedicated one room, which is usually used as library and reading room, to the data collection. The area of the room was about 60 m^2. Half of the walls were covered by bookshelves. The other half was plain with some decorations attached to the walls. Furniture, such as tables, chairs, and couches reduced the reverberation in the room. Recordings were performed with the new data collection setup (cf. Figure 5.3). In total, data of 126 children (68 female and 58 male) at the age of 8.6 ± 1.1 years were gathered in April 2007. The dataset is called *Hannover-07* subsequently.

5.2.4 Karlsruhe (Baden)

For the representation of a south-western dialect, Karlsruhe in Baden was chosen. Three schools — the Gutenberg Schule, the Nebenius-Grundschule, and the Hans-Thoma Schule — agreed to contribute to the data collection which was performed between May and June 2007. Recording conditions were similar in all three schools since all of them provided individual rooms for the data collection. In the Gutenberg Schule, a room which is usually dedicated to medical examinations was used. It was a rather small room with an area of about 25 m^2. The room was calm and did not have any noise sources. In May 2007, 63 children (30 female and 33 male) at an average age of 8.3 ± 1.2 years were recorded. The dataset is referred to as *Gutenberg-07*.

In the Nebenius-Grundschule, the data collection was performed in the library of the school. The room had an area of about 50 m^2. On the walls were some bookshelves which reduced reverberation. The 44 children (25 female and 19 male) at the age of 8.7 ± 0.9 years are labeled as *Nebenius-07*.

The recordings in the Hans-Thoma Schule were performed in the office of a teacher. The area of the office was about 20 m^2. The office was rather quiet although a classroom was situated right next to it. In this school only one second class with 24 children (12 female and 12 male) was recorded. The group of children with an average age of 7.7 ± 0.6 is called *Hans-Thoma-07* in the following. The union of all three schools in Karlsruhe is referred to as *Karlsruhe-07*.

5.2.5 Leipzig (Saxony)

To represent a dialect of eastern Germany, Leipzig was chosen. The Dritte Grundschule supported this work by admission of the recordings. They dedicated one office in the cellar of the school to the recordings. The area of the room was about 15 m^2. Overall recording situation was calm and neither echo or reverberation posed a problem to the sound quality. Between March and July 2007, 61 children (40 female and 21 male) with an average age of 7.9 ± 2.1 years were collected. This dataset is called *Leipzig-07* throughout this work.

5.3 Patient Groups

Collection of the patient groups started already before the development of the first version of PEAKS. Therefore, this section is divided into two parts. First, the data which were gathered before the first version of the PEAKS client was finished are described. Next, the patient data recorded with the PEAKS client are presented in detail.

5.3.1 Preliminary Recordings

Recording of children with cleft lip and palate started already in 2002 in the Oral and Maxillofacial Clinic of the University Hospital of Erlangen. The hospital has an examination room for the follow-up care of the children with CLP. The recordings for this work were gathered during the regular out-patient examination. The room in which the recordings took place has an area of about 15 m^2. The room contains two shelfs, a patient chair, and an endoscope with PC tower. In the room no echo nor reverberation was audible. Additional noise sources, like the hardware required for endoscopy, were always turned off before any audio data were captured. Until late January 2006, before the first version of the PEAKS client became available, 123 children (54 female and 69 male) were recorded. All data of the CLP children recorded with the old recording software are referred to as *CLP-02-06*. The children had an average age of 8.2 ± 3.6 years. 16 of the children had an isolated cleft lip, 30 an isolated cleft palate. In the children with cleft lip and palate, 57 had the unilateral phenotype while 20 had the bilateral one. The recording was performed directly at the PC, however, there was no automatic segmentation performed during the test. As microphone a dnt Call 4U Comfort at a sampling frequency of 16 kHz quantized with 16 bit was used. The children wore it as head set. Then, the recording was started. In this first version of the recording software, the PLAKSS slides had to be shown on paperback to the children. The PC was just used to store the audio data. This procedure had several disadvantages. In order to process the data with speech recognition technology, at least some segmentation is desired. During spontaneous speech, some pauses occur at phrase or sentence boundaries. These were used to automatically segment the data into turns. In total, the data were segmented into 5176 turns, i.e., 42 turns per child on average. Before any evaluation could be performed all of the data had to be transliterated. This is a laborious process which has a real-time factor of about 10. The transliteration contains 17831 words, i.e. 3.4 words per turn. All the data had to be listened to manually, and every word uttered had to be noted down. In total, the vocabulary contains 605 words plus 1062 word fragments and pathologic word alternatives. Compared to the 99 target words, 605 words seem a lot more. However, one has to keep in mind that all words which were uttered by the children were written down in the transliteration. It contains also carrier sentences like “this is a …” or “I can see a …”. Moreover, description words, such as colors and other item properties appear in the transcript. A serious difficulty in transliteration of pathologic speech data are the word fragments which also contain pathologic word alternatives. For an inexperienced listener, it is very

category	criterion	abbreviation
nasality	hypernasality	HN
	nasalized consonant	NC
backing	laryngeal replaced	LR
	pharyngeal backing	PB
articulation	palatalization	PA
	weakened plosives	WP
lisp	lateralization	LA
	interdentalization	IN

Table 5.1: Overview on the pronunciation rating criteria

difficult on the one hand to understand the child at all and on the other hand to find a good alphabetic representation of the word.

The first 31 recordings (9 girls and 22 boys) of the *CLP-02-06* dataset were perceptively assessed by five speech professionals using PEAKS. A screen shot of the evaluation screen is shown in Appendix A.3 (cf. Figure A.6). The children with CLP were in the age from 4 to 16 years (mean 10.1±3.8 years) at the time of the recording. Two had an isolated cleft lip, five an isolated cleft palate, 20 a unilateral cleft lip and palate, and four a bilateral cleft lip and palate. The total duration of the children's audio files was 120 minutes, consisting of 5330 words in 2209 turns. The vocabulary contains the 795 words which occur in the data (97 unique words of the test, 266 additional adjectives and nouns which were used by the children to explain the pictures, and 432 additional representing word fragments). The average turn length is short with 2.4 words. The recordings showed a wide range in intelligibility. This subset of *CLP-02-06* is called *CLP-Intel*. It is used for intelligibility assessment in Chapter 6.

A subset of 26 children (5 female and 21 male) of *CLP-Intel* was furthermore assessed by a speech therapist using the forms presented in Appendix A.2. The speech therapist had been working with children with cleft lip and palate for many years. Therefore, she could differentiate many criteria: "hypernasality", "nasalized consonant", laryngeal backing as "laryngeal replaced" and "pharyngeal", "palatalization", "weakened plosives", "lateralization", and "interdentalization". The criteria listed in Table 5.1 were selected in order to match Table 2.2. Two of the children in the dataset had an isolated cleft lip, three an isolated cleft palate, 19 unilateral CLP, and another two bilateral CLP. Assessment was only performed on the PLAKSS target words. Hence, only the 1916 words of the transliteration which could be mapped onto one of the 99 target words were used. In average, 73.7 target words could be recovered from the audio data per child. Some of the words could not be obtained from the audio files because they were either not uttered by the child, or the uttered word could not be mapped to the target word, i.e., an uncommon word alternative or word fragment was uttered by the child. Annotation according to the forms in Appendix A.2 based on the target phonemes of the PLAKSS words. This is basically a phoneme level annotation. However, only some phonemes of the words are marked. In order to obtain a full phoneme level annotation, a second perceptive assessment by a speech expert was performed based on the annotations of the first expert. Using the phoneme level annotation module of PEAKS (cf. Appendix A.3), the evaluation

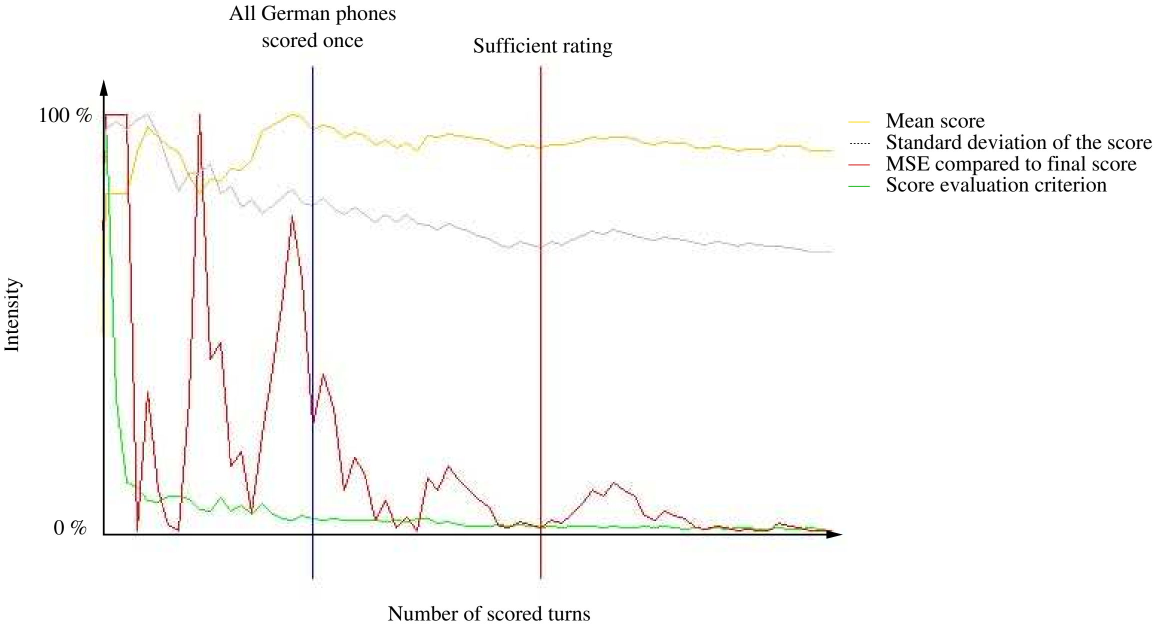


Figure 5.4: If the turns are in a sequence which present all German phonemes early the development of the mean value and the variance of the score gets stable. Hence, a criterion can be defined which allows to abort the assessment procedure if the mean value of the score and its standard deviation get stable.

was performed directly at the PC. With these detailed annotations, frame, phoneme, word, and speaker level experiments can be performed on this dataset. Hence, the set is called *CLP-Phone-Eval*.

In 2004, a preschool study of children's speech was performed at the Department of Phoniatrics and Pedaudiology of the University Hospital Erlangen. All children were tested with the PLAKSS test and recorded with a DAT recorder at 48 kHz and 16 bit quantization. None of the children had cleft lip or palate. However, some of the children had other language and speech disorders. Hence, the audio data could not be used for this work. Transliteration, however, of the data showed to be beneficial. The transliteration of the data could be employed to extend the list of frequent word alternatives of the target words of the PLAKSS test. Moreover, the data could be used to train the language models required in some segmentation steps. The dataset contains speech of 50 children (14 female and 36 male) at an average age of 5.3 ± 1.1 years and is referred to as *Preschool*.

5.3.2 Recordings with PEAKS

The first patient was recorded with PEAKS on January the 20th, 2006. Subsequently, until March the 17th, 35 children (13 female and 22 male) with CLP at the age of 8.5 ± 3.5 years were recorded to form the first patient group gathered with PEAKS — *CLP-Intel2*. The recordings were performed with the same microphone as the previous patient groups (dnt Call 4U Comfort) at 16 kHz with 16 bit quantization in the same room. Eight of the children had an isolated cleft lip, nine an isolated cleft palate, 13 unilateral CLP, and five bilateral CLP. Due to the new data collection procedure, the exact time is known when the supervisor of the test moved from one slide to another. Hence, manual transcription and segmentation were no longer required since the text reference of the respective slide can be used instead.

All data of *CLP-Intel2* were also evaluated by five speech professionals using the evaluation module as shown in Appendix A.3 (cf. Figure A.6). The evaluation was performed as discussed in Chapter 3.1. Since the speech experts complained about the laborious evaluation procedure of every single turn, a faster procedure was designed. Using the data of the first evaluation session on *CLP-Intel*, it could be shown that already about 30 % of the scores would be sufficient. However, the turns have to be reordered since the setup of the PLAKSS test focuses on different phonemes in the beginning than in the end. Otherwise it might happen that some phonemes did not appear although already half of the test was scored. In some children, however, only certain phonemes are affected by the clefting. Hence, the slides were brought to a new sequence which presents all phonemes of the German language already in the first third of the audio data. Figure 5.4 shows an example. In the beginning of the scoring procedure, the mean value and the standard deviation of the score strongly vary. This also causes a high mean square error compared to the final score of the patient. At the time when all phonemes appearing in the speech data of the patient were listened to at least once, the error drops. Moreover, the mean value and the standard deviation of the score show only little variation until the end of the assessment session. Further scores contribute only little to the final score. The heuristic criterion $e_{\text{crit}}(t_r)$ to decide whether to continue the assessment at turn t_r or not is computed as follows:

$$\begin{aligned} e_{\text{crit}}(t_r) &= \left\{[\mu(t_r) - \mu(t_r - 1)]^2 + [\sigma(t_r) - \sigma(t_r - 1)]^2\right\} \\ &\quad +0.5\left\{[\mu(t_r - 1) - \mu(t_r - 2)]^2 + [\sigma(t_r - 1) - \sigma(t_r - 2)]^2\right\} \\ &\quad +0.2\left\{[\mu(t_r - 2) - \mu(t_r - 3)]^2 + [\sigma(t_r - 2) - \sigma(t_r - 3)]^2\right\} \end{aligned} \tag{5.1}$$

$\mu(t_r)$ denotes the mean score and $\sigma(t_r)$ the standard deviation computed only with the ratings known at time t_r. The rating was considered to be sufficient when at least half of the turns were scored and the criterion $e_{\text{crit}}(t_r)$ was less than 10 % of the maximal variation $v_{\max}$ until turn t_r. $v_{\max}$ is computed as

$$v_{\max} = \underset{t_r}{\operatorname{argmax}}\, [\mu(t_r) - \mu(t_r - 1)]^2 + [\sigma(t_r) - \sigma(t_r - 1)]^2 . \tag{5.2}$$

This procedure showed to have about half of the error compared to other heuristic procedures as shown in Chapter 6. In this manner about 30 % of the ratings were omitted.

Until 2008, more and more data of patients were gathered. In January 2008, a total of 189 children (85 female and 104 male) were already recorded. 21 of them had a cleft lip, 48 a cleft palate, 102 unilateral CLP, and 18 bilateral CLP. The average age at the time of the recording was 9.2 ± 4.3 years. The dataset is called *CLP-06-08* in the following.

Most of the control data were collected from children in elementary school age. Hence, a subset of *CLP-06-08* was selected to form an age-matched patient group. All children in *CLP-06-08* in the age from 6 to 10 years were included to form the set

CLP-School. It contains speech of 59 children (27 female and 32 male) at the age of 8.5 ± 1.6 years. 14 of the children had an isolated cleft lip, 16 an isolated cleft palate, 24 unilateral CLP, and five bilateral CLP.

A subset of *CLP-School* was evaluated on phoneme level using the phoneme level annotation module of PEAKS (cf. Appendix A.3). All phonemes of the PLAKSS test target words were assessed by two experienced speech therapists. As described in the literature review in Chapter 2, the most important and distinct feature in speech of children with CLP is the enhanced nasal air emission. Therefore, the speech therapists investigated only the nasality of the speech data of the subset. Since the speech therapists were less experienced than the speech expert who annotated the data in the *CLP-Phone-Eval* data, the categories "hypernasalization" and "nasalized consonants" were merged into the category "nasalization". The resulting set *CLP-Phone-Eval2* contains speech data of 32 children (17 female and 15 male). Their average age was 8.7 ± 1.7 years at the time of the recording. Five of the children had a cleft lip, seven a cleft palate and 20 a unilateral cleft lip and palate. Children with bilateral CLP did not appear in the data set.

A subset of *CLP-06-08* with twelve children was recorded two times. At the time of the first recording, the children were 9.4 ± 3.9 years old on average. The group contained six male and and six female children. Nine of the children had an isolated cleft lip and palate, one a bilateral cleft lip and palate, one a cleft lip, and one a cleft palate. The second recording was performed one year later. This means that the progress the children made can be measured in this group and is therefore called *CLP-Progression*.

5.4 Training Data of the Speech Recognition System

The speech recognition system had been trained with acoustic information from spontaneous dialogues of the VERBMOBIL project [Wahls 00] and normal children's speech.

The training population of the VERBMOBIL project consisted of normal adult speakers from all over Germany and thus covered all regions of dialect. All speakers were asked to speak "standard" German. 90 % of the training population (85 male and 47 female) were younger than 40 years. The used subset had a total duration of 4.4 hours and is also referred to as "VERBMOBIL tiny" [Hader 02, p.39].

The speech data of non-pathologic children (23 male and 30 female) were recorded at two local schools in Erlangen — the Montessori Schule and the Ohm Gymnasium — and consisted of read texts. The 25 children (17 female and 8 male) of the Montessori Schule were in average 11.6 ± 0.7 years old. 10.5 ± 0.5 years is the average age of the 18 children (12 female and 16 male) recorded at the Ohm Gymnasium. The mean age of all children in both groups together is 11.0 ± 0.8 years. As microphone the same dnt Call 4U Comfort was employed as for the collection of the patient groups in the Oral and Maxillofacial Clinic of the University Hospital of Erlangen. Sampling rate was 16 kHz, quantization 16 bit. The children are a subset of the children who were recorded for the Aibo database [Batli 04]. Moreover, non-native English speech data were collected from the same children. Results of the evaluation are presented in [Hacke 07a]. The total playing time of the children's speech was 9.1 hours.

label	location	#	recording date	avg. age	chapter
Erna-Zink-07	Erlangen	38	March 2007	5.7 ± 0.7	5.2.1
Michael-Poeschke-06	Erlangen	89	January 2006	8.8 ± 1.3	5.2.1
Michael-Poeschke-07	Erlangen	76	April 2007	8.5 ± 1.4	5.2.1
Michael-Poeschke-08	Erlangen	157	February 2008	8.4 ± 1.2	5.2.1
Sabel-07	Nuremberg	48	March 2007	13.2 ± 1.2	5.2.2
Hannover-07	Hannover	126	April 2007	8.6 ± 1.1	5.2.3
Gutenberg-07	Karlsruhe	63	May 2007	8.3 ± 1.2	5.2.4
Nebenius-07	Karlsruhe	44	May 2007	8.7 ± 0.9	5.2.4
Hans-Thoma-07	Karlsruhe	24	June 2007	7.7 ± 0.6	5.2.4
Karlsruhe-07	Karlsruhe	131	May 2007	8.3 ± 1.1	5.2.4
Leipzig-07	Leipzig	61	March 2007	7.9 ± 2.1	5.2.5

Table 5.2: Summary of the data collected as control groups: In total, 726 children were recorded as controls.

All adult speaker's data were then vocal tract length normalized (cf. Chapter 4.2.5) to simulate children's speech. The scaling factor was determined experimentally on the children's evaluation set. During training an evaluation set was used that only contained children's speech. Optimal results were obtained at a scaling factor of 0.83.

5.5 Summary

In this chapter the data acquisition for this work was described. All data were collected using the PLAKSS test. The test contains all German phonemes in different positions. The target words are shown to the children as pictograms. Hence, the test is also suitable for children who are not yet able to read.

Control groups were collected in cities all over Germany to test for dialectal influences. In the first year, data was collected with a complex setup in order to optimize the sound quality. In the subsequent years, more data were gathered with a more simple setup: In order to guarantee a similar audio quality on different recording computers and on different sites, a USB head set was used. This simplified the setup a lot. A summary of the recorded control data is presented in Table 5.2.

Since 2002, 312 children with cleft lip and palate were recorded in the Oral and Maxillofacial Clinic of the University Hospital of Erlangen. Until 2006, 123 children were recorded directly to a PC located in the clinic, i.e., the predecessor of PEAKS. Hence, the time when the supervisor moved to the next slide was unknown. All data had to be transliterated manually for these children. A subset of 31 children were perceptively scored with respect to their speech intelligibility. 26 of these children were further assessed on phoneme level. In order to reduce evaluation efforts PEAKS was introduced in January 2006. A total of 189 children were recorded until 2008. The intelligibility of 35 of them was assessed by five speech experts. 59 of these children were in elementary school age at the time of the recording. Phoneme level evaluation was performed on a subset of 13 children. Table 5.3 gives a summary of the patient data collected for this work.

label	# CL	# CP	# UCLP	# BCLP	$\sum$	avg. age	chap.
CLP-02-06	16	30	57	20	123	8.2 ± 3.6	5.3.1
CLP-Intel	2	5	20	4	31	10.1 ± 3.8	5.3.1
CLP-Phone-Eval	2	3	19	2	26	9.4 ± 3.3	5.3.1
Preschool	-	-	-	-	50	5.3 ± 1.1	5.3.1
CLP-06-08	21	48	102	18	189	9.2 ± 4.3	5.3.2
CLP-Intel2	8	9	13	5	35	8.5 ± 3.5	5.3.2
CLP-School	14	16	24	5	59	8.5 ± 1.6	5.3.2
CLP-Phone-Eval2	5	7	20	-	32	8.7 ± 1.7	5.3.2
CLP-Progression	1	1	9	1	12	9.4 ± 3.9	5.3.2

Table 5.3: Overview on the patient data collected in Erlangen: 312 children with cleft lip and palate were recorded. Note that some of the listed sets are subsets of larger sets.

At the end of this chapter, the data used for the recognizer training was described. The data consist of speech of children from two local schools and adults' speech from the VERBMOBIL project which was adapted to children's speech.

Chapter 6

Experiments

In this chapter the previously described methods are evaluated to whether they are suitable for clinical purposes. Therefore, the algorithms have to be as reliable or even more reliable as human raters since they are regarded the reference. The evaluation of one expert is still subjective. Therefore, the perceptive evaluation of multiple speech professionals is desirable. Hence, the results concerning the perceptive evaluations of the databases are presented in the first section of this chapter.

Next, the results on the speech intelligibility are described. First experiments are based on the fully transliterated patient data, as described in Chapter 5.3.1. This is followed by results which are computed automatically. In order to enhance the prediction quality further, prosody is incorporated into the assessment procedure. Finally, the results obtained on the control groups are presented.

Intelligibility is a global outcome parameter. A more detailed analysis is possible if single articulation disorders of the patients' speech are evaluated. Hence, the third section of this chapter deals with the automatic assessment of the patients' articulation. Again, first the results on the transliterated data are presented and discussed. Then, the results on the non-transliterated data are presented.

The last section of this chapter presents the results of the visualization using the extended Sammon mapping. With this technique it is possible to remove the differences between the different acoustic conditions to create a unified map for all microphones.

6.1 Perceptive Evaluations

As described in Chapter 3, two different kinds of perceptive evaluations were performed for this work: Intelligibility assessment using Likert-scales [Liker 32] and phoneme level annotation of aspects according to Chapter 2.

6.1.1 Perceptive Scoring of the Intelligibility

As standard procedure for the perceptive evaluation of the speech intelligibility a five-point Likert-scale was chosen for this work. As shown in [Maier 07d], such a procedure converges already with four expert listeners. Hence, four to five experts were chosen.

CLP-Intel		
rater	mean of other raters	
	r	ρ
rater B	0.95	0.92
rater K	0.94	0.93
rater L	0.94	0.93
rater S	0.94	0.92
rater W	0.96	0.92

Table 6.1: Agreement of the raters on the *CLP-Intel* database measured with Pearson's correlation r and Spearman's correlation ρ

criterion	error / patient	# of scores	error / score
"three turns"	0.385	93	0.128
"five turns"	0.227	155	0.045
"ten turns"	0.091	310	0.009
"30 turns"	0.028	930	0.001
"every phoneme at least once"	0.036	714	0.002
"e_{crit}" (cf. Chapter 5.3.2)	0.024	1124	0.001
"every phoneme at least once" & e_{crit}	0.017	1095	0.000

Table 6.2: Comparison of different criteria for the abbreviation of the scoring procedure, computed using the 2209 turns of the *CLP-Intel* database.

The first dataset on which the intelligibility was scored, is *CLP-Intel*. To ensure convergence, a number of five experts was chosen. Table 6.1 shows the inter-rater correlations (cf. Chapter 4.2.1) of the speech intelligibility scores on the dataset. The correlations denoted in the table are computed between one rater and the mean of the other raters. Both Spearman's and Pearson's correlation coefficients are in the same range. The overall agreement is very high with significant correlations between 0.92 and 0.96 ($p < 0.001$). Alpha (cf. Chapter 4.2.1) was also very high with 0.75. Since the weighted multi-rater Kappa (cf. Chapter 4.2.1) is only defined on integer values the average scores had to be rounded to the next integer for its computation. Kappa also showed high agreement with 0.59.

The labeling of each turn individually was perceived as laborious and monotonic by most of the labelers. A method for the abbreviation of the rating procedure was desired. Several methods were investigated. As criteria the evaluation of a fixed number of turns and the heuristic error criterion as defined in Chapter 5.3.2 were investigated. Table 6.2 gives an overview on the errors obtained with the different criteria. The first four criteria concern only a fixed number of turns which have to be evaluated per patient. With an increasing number, the error per patient and per score is reduced. However, there might be better criteria to determine whether the number of scores is already sufficient or not. Hence, the criteria "every phoneme at least once" and e_{crit} (cf. Eq. 5.1) were investigated. Both turn out to produce very

CLP-Intel2		
rater	mean of other raters	
	r	ρ
rater M	0.92	0.88
rater L	0.93	0.88
rater S	0.95	0.92
rater W	0.90	0.87

Table 6.3: Agreement of the raters on the *CLP-Intel2* database

	NC	LR	PB	WP	LA	IN	Intel
HN	0.52 (**)	-0.07	-0.08	0.73 (**)	0.03	-0.11	0.48 (*)
NC		0.57 (**)	0.43 (*)	0.55 (**)	-0.10	-0.08	0.73 (**)
LR			0.87 (**)	0.31	-0.04	-0.12	0.60 (**)
PB				0.22	0.07	-0.19	0.55 (**)
WP					-0.08	0.00	0.59 (**)
LA						-0.20	0.22
IN							-0.13

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

Table 6.4: Pearson's correlation between the different evaluation criteria: hypernasality (HN), nasalized consonants (NC), laryngeal replacing (LR), pharyngeal backing (PB), weakened plosives (WP), lateralization (LA), interdentalization (IN), and the intelligibility (Intel).

low errors as well. The best, i.e. lowest, error was found as a combination of both methods. The error found per score is close to zero.

On the second database — *CLP-Intel2* — assessment of the intelligibility was only performed by four experts. Furthermore, the scoring procedure was abbreviated as previously described. Again, the agreement was very high with an Alpha of 0.80 and a Kappa of 0.68. Table 6.3 presents Pearson's and Spearman's correlation coefficients. The correlations are in the same range as on the *CLP-Intel* database although slightly lower. All correlations are significant ($p<0.001$), and the consistency is very good.

6.1.2 Perceptive Assessment of the Altered Articulation

The phoneme level evaluation of the children's speech was performed on two databases. While the data of *CLP-Phone-Eval* were completely transliterated, no manual transliteration on the *CLP-Phone-Eval2* data was performed. Annotation was done for typical alterations such as "hypernasality" (HN), "nasalized consonant" (NC), laryngeal backing as "laryngeal replacing" (LR) and "pharyngeal backing" (PB), "weakened plosives" (WP), "lateralization" (LA), and "interdentalization" (IN) (cf. Table 5.1) by an experienced speech therapist on the *CLP-Phone-Eval* data. "Palatalization" was not found in the data by the speech therapist. The *CLP-Phone-Eval2* data were evalu-

	Component		
	1	2	3
HN	-0.17	**0.94**	0.11
NC	**0.54**	**0.68**	-0.06
LR	**0.97**	0.11	-0.01
PB	**0.94**	0.02	0.14
WP	0.20	**0.88**	-0.07
LA	-0.06	-0.07	**0.79**
IN	-0.15	-0.07	**-0.75**

Table 6.5: Factor loadings on the three principal factors of the perceptive evaluation of *CLP-Intel*

ated by two speech therapists according to the criterion “nasality”, i.e., the union of “hypernasality” (HN) and “nasalized consonant” (NC).

In total, 867 words in the *CLP-Phone-Eval* were identified to be part of one of these categories. A detailed overview on the patients and the number of marked words is shown in Appendix A (cf. Table A.7). Since the total number of words differed from child to child, the percentage of marked words was used for all speaker level computations. On speaker level, Table 6.4 shows highly significant correlations between HN, NC, WP, and the mean of intelligibility scores given by the five speech experts — denoted as “Intel”. Furthermore, LR is highly correlated with PB ($r = 0.87; p < 0.001$), i.e. both types of backing are closely related. Most of the criteria are related to the speech intelligibility. Only LA and IN show no significant correlation to the intelligibility.

In order to gain further insight on the data factor analysis was performed with PCA (cf. Chapter 4.2.3). Since intelligibility is a global criterion it was excluded from the factor analysis. Using scree analysis [Catte 66], three factors were extracted (cf. Figure A.8). The factor loadings are presented in Table 6.5. The first component shows the highest loadings on the criteria LR and PB. A slight loading of NC is also found. Hence, the component could be interpreted as a “backing” component. Component 1 has a correlation of $r = 0.55$ ($p < 0.01$) with the intelligibility scores of the experts. The highest loadings of component 2 appear in HN, WP, and NC. Thus, this component can be regarded as the “nasalization” component. It also has a rather high correlation to the speech intelligibility ($r = 0.58; p < 0.01$). In the third component, only LA and IN have high loadings. Note that both loadings have the opposite sign, i.e. they point into diametric directions. The third component, hence, can be regarded as a “lisp” component. No significant correlation to the speech intelligibility is found ($r = 0.20$). Linear regression of all three components to the speech intelligibility shows a significant correlation of $R = 0.82$ ($p < 0.001$). However, the contribution of the third component is insignificant with $p = 0.12$. This means that the global outcome parameter of the test — the speech intelligibility — can be divided into a backing component and into a nasalization component in CLP speech. A third component is found — the lisp component. However, it is not or only slightly related to the speech intelligibility.

nasality	nasal (rater 1)	non-nasal (rater 1)
nasal (rater 2)	127	203
non-nasal (rater 2)	152	2499

Table 6.6: Confusion between both speech therapists who rated the criterion "nasality" on the *CLP-Phone-Eval2* database

On the *CLP-Phone-Eval2* data, only the criterion "nasality" was scored. The agreement is quite good. Table 6.6 shows that the agreement on non-nasality is very high. 2499 of the 2981 words were not marked as "nasal". However, only 127 words were marked as "nasal" by both raters. This corresponds to a true positive rate of human rater 1 of 45.5 % at a false positive rate of 7.5 %. Rater 2 had a true positive rate of 61.5 % with a false negative rate of 5.7 %. Hence, both raters know "non-nasal" well, but their agreement on "nasal" is only moderate.

Correlation on speaker level also showed good consistency. When the percentages of marked words per speaker of both raters are compared, a correlation of 0.80 is obtained.

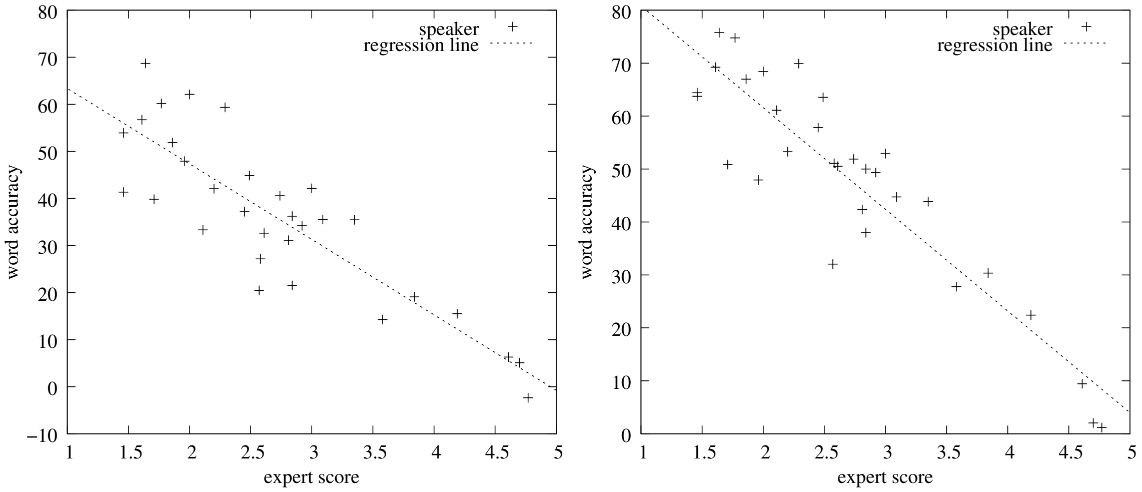
6.2 Automatic Evaluation of the Intelligibility

Now that the agreement of the human raters was discussed, their evaluations can be used to create a reference for an automatic speech assessment system. The first semi-automatic and automatic experiments focused on the evaluation of the speech intelligibility of the children. Experiments concerning the intelligibility were performed on the *CLP-Intel* and the *CLP-Intel2* databases. On the *CLP-Intel* data, all speech had to be transliterated before the processing. Hence, the assessment is just semi-automatic. Fully automatic assessment was performed on the *CLP-Intel2* database.

6.2.1 Semi-Automatic Evaluation

For the *CLP-Intel* data, a complete transliteration was available. Because of the very limited amount of data, we used the transliteration of all recordings as training set for the unigram language model. The vocabulary size is still large enough so that the acoustic realization by the children has high enough an influence on the word accuracy (the test set perplexity is 94).

Since children's speech is more difficult for speech recognition, the use of adaptation for the improvement of the speech recognition system was first investigated. When compared to the average of the raters, the word accuracy (WA; cf. Chapter 4.2.4) for the recognizer had a correlation of $r = -0.89$ and $\rho = -0.85$ for the adapted case and $r = -0.84$ and $\rho = -0.80$ for the non-adapted case. Note that these values are slightly different from [Maier 06b] since two additional raters became available for the data. Figure 6.1(a) shows the correspondence between the non-adapted recognizer and the expert panel while Figure 6.1(b) displays the adapted case. The coefficients are negative because high recognition rates come from "good" speech with a low score number and vice versa (note the regression line in the figure). Because the



(a) non-adapted recognizer

(b) adapted recognizer

Figure 6.1: Comparison between adapted and non-adapted speech recognizers for the measurement of the speech intelligibility

agreement in the adapted case was better, speaker adaptation is always performed before the decoding step for the rest of this work.

Table 6.7 lists the correlations between the individual raters and the adapted speech recognizer. The correlations are all in the same range. The highest correlations are found between the mean opinion of the experts and the speech recognition system.

If the adapted speech recognizer is added to the group of raters, problems for the computation of Kappa and Alpha occur. For both, the scaling of the recognizer's scores has to be adjusted to the dimensions of the Likert-scales. This can be performed in a linear, equidistant way or non-linearly with different interval sizes. With the spacing as proposed in [Schus 06a] (borders at 0, 15, 25, and 40 % WA), a Kappa of 0.53 and Alpha of 0.75 are obtained. Equidistant borders yield a Kappa of 0.44 and Alpha of 0.58. If the borders are chosen to be optimal according to the given criterion (borders at -5.2, -3.9, 17.5, and 40.9 for Kappa and -5.2, 16.1, 27.5, and 48.9 for Alpha), Kappa yields 0.56 and Alpha 0.77. This means that for all comparisons with the Multi-Rater-Kappa and Alpha between the human raters and the automatic scoring, such a score transformation has to be applied. Hence, all scores computed with Kappa and Alpha are highly dependent on this transformation and show a lot of variation. Regarding Alpha, the first result with the scaling as presented in [Schus 06a] is the same as in the group of experts. Using the equidistant spacing, the values of Kappa and Alpha are quite a lot lower than in the experts' group. The optimally computed Alpha turns out to be even higher than in the experts without the recognizer. Pearson's and Spearman's correlation coefficients allow for a comparison between the experts and the raters without any of these ambiguities. Hence, further computations of Kappa and Alpha are skipped.

	WA		WR	
	r	ρ	r	ρ
rater B	-0.87	-0.86	-0.86	-0.85
rater K	-0.84	-0.82	-0.83	-0.80
rater L	-0.86	-0.81	-0.84	-0.76
rater S	-0.86	-0.82	-0.86	-0.82
rater W	-0.85	-0.81	-0.83	-0.76
all raters	-0.89	-0.85	-0.88	-0.82

Table 6.7: Correlations between the different raters and the recognition rate of the adapted speech recognizer in the *CLP-Intel* data

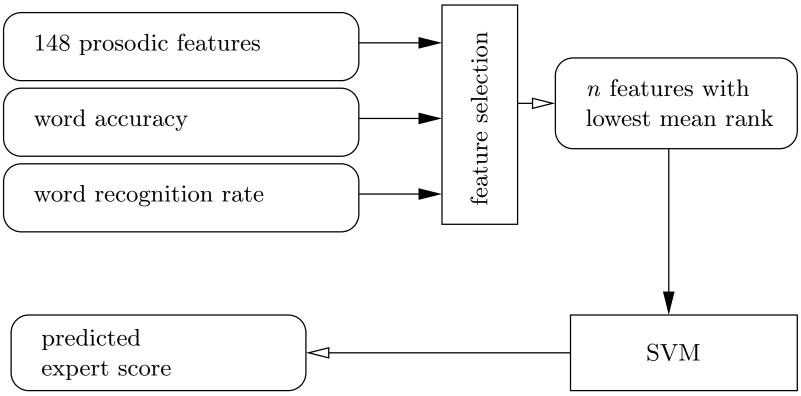


Figure 6.2: Proposed system for the prediction of the expert scores

feature	prediction SVR		reference raters
	r	ρ	
word accuracy	0.86	0.84	all raters
+ minimum EnergyRegCoeffWord	0.86	0.82	all raters
+ mean Mean_shimmer	**0.87**	0.82	all raters
+ minimum F0MeanWord	0.85	**0.87**	all raters
word accuracy	**0.85**	**0.83**	rater B
word accuracy	0.82	0.80	rater K
+ minimum F0MaxWord	**0.84**	**0.86**	rater K
word accuracy	**0.83**	**0.78**	rater L
word recognition rate	**0.82**	**0.79**	rater S
word accuracy	**0.84**	**0.81**	rater W

Table 6.8: Prediction of the experts' scores by different feature sets on the *CLP-Intel* database

In [Hader 06a] it was shown that — besides the recognition rate of a speech recognizer — prosodic features also hold information on the intelligibility. Hence, a system was designed to include several information sources into the prediction process. This automatic evaluation system employs SVR (cf. Chapter 4.2.2) for the prediction of the experts' scores.

As displayed in Figure 6.2, the system utilizes on the one hand the WA and the WR of a speech recognizer (cf. Chapter 4.2.4). On the other hand, 148 prosodic features (cf. Chapter 4.2.4) are used in the system. So 150 features are obtained in total. In order to select a subset of the features, we applied the Maximum R algorithm based on the multiple regression/correlation analysis as described in Chapter 4.2.3. The algorithm builds — based on the best $(n-1)$ subset — all possible sets with n features and picks the set with the best regression to the target value (here: the mean opinion of the experts). The algorithm returned better features than other feature selection algorithms, like correlation-based feature subset selection [Hall 98] or consistency subset evaluation [Liu 96]. However, the algorithm can select $m-1$ features at most where m is the number of subjects in the test set. If a feature was not selected, we assigned rank 149.

All evaluations presented here were done in a leave-one-speaker-out (LOO) manner since the number of patients in each group is rather small. In order to present a feature ranking for the feature selection, we computed the mean rank of all LOO iterations for each feature. This, however, does not mean that the particular feature has been selected in all LOO iterations.

The combination of the prosodic features and the result of the speech recognizer is beneficial for the prediction of experts' scores (cf. Table 6.8). Now, all correlations are positive since they denote the dependency between the predicted score and the actual score. The best feature for the prediction of the intelligibility is for all raters either the WA or the WR. The mean of all raters is best predicted in the sense of Pearson's correlation with the word accuracy, the minimum energy contour regression slope per word, and the mean of the mean shimmer in each turn. In terms of Spearman's correlation, the prediction can be improved by adding the minimum mean F_0 per

	WA		WR	
	r	ρ	r	ρ
rater M	-0.83	-0.78 (*)	-0.88	-0.90 (*)
rater L	-0.78	-0.75	-0.84	-0.86
rater S	-0.76	-0.72 (*)	-0.85	-0.88 (*)
rater W	-0.78	-0.76	-0.86	-0.83
all raters	-0.82	-0.81 (*)	-0.90	-0.93 (*)

Table 6.9: Correlations between the different raters and the recognition rate of the adapted speech recognizer in the *CLP-Intel2* data: Significant differences ($p < 0.05$) between WA and WR are marked with (*).

word. As shown in Table 6.8, the selection of the first feature does not yield any improvement. The combination with more features, however, refines the prediction of the experts' scores.

For the prediction of individual experts, only the prediction of a single expert could be improved (rater K) by adding one prosodic feature. In general, the prediction of the individual raters is performed with a Pearson correlation $r > 0.80$ and a Spearman correlation $\rho > 0.75$. Note that the correlations are lower than in the previous experiment since all experiments were conducted in a LOO manner. $|r| = 0.89$ of the first experiment is reduced to 0.86, and $|\rho| = 0.85$ becomes 0.84 if the experiments are computed in LOO mode.

6.2.2 Fully Automatic Evaluation

Until this point the transliteration of all audio data was a necessity which could not be omitted. However, this procedure costs a lot of time and manpower. Since a new recording and evaluation software was developed, the exact time when the reference slide was moved to the next slide is known in the *CLP-Intel2* data. This information can be exploited to approximate a reference word chain. This reference word chain contains just the words which are shown on the slide. Unfortunately, this is not sufficient to calculate a good word accuracy since most of the children use carrier sentences like "This is a ..." which are regarded as wrongly inserted words even if the recognition were perfect. Table 6.9 shows the effect of these carrier words: The correlation between the human evaluation and the WA is lower than the correlation between the human experts and the WR in all cases. Cases in which the difference is significant are marked with (*). The best Spearman correlation is found with $\rho = -0.93$ which is better than all human experts (cf. Figure 6.3).

Additional use of prosody does not improve the correlation further (cf. Table 6.10). If the mean of all raters is the target value, only the WR is selected as best feature with a Pearson correlation of 0.89 and a Spearman correlation of 0.92 in LOO evaluation. For half of the individual raters, however, the use of prosody yields an additional improvement. The correlation to rater M is increased from 0.86 to 0.88 and from 0.89 to 0.90, respectively. For rater S the agreement to the automatic system is enhanced from 0.79 to 0.84 in terms of Pearson's correlation and from 0.85 to 0.89 for the case of Spearman's correlation.

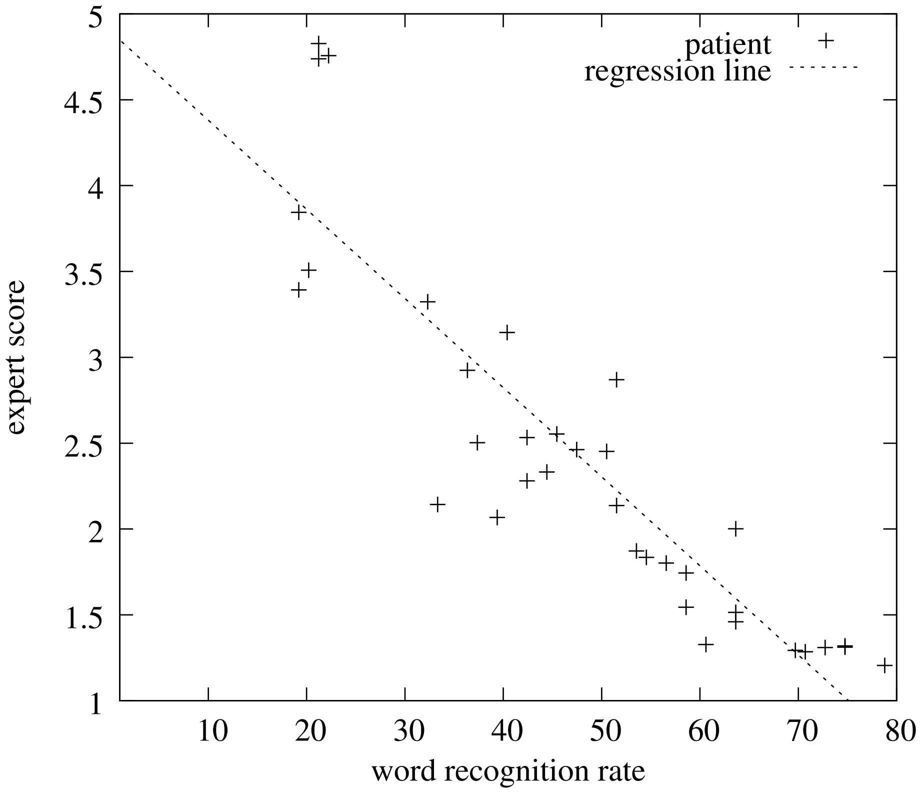


Figure 6.3: Agreement between the mean scores of the panel of experts and the WR: The evaluation of the speech recognizer is as reliable as the best human raters.

We relate the weak performance of the additional prosodic features to three different reasons. First of all, the PLAKSS test is based on single words. Hence, there is only little prosody is to be expected. Many of the children just named the single items. Only some of the children create connected sentences to describe the items, i.e., there is not much prosody in the data at all. Secondly, the quality of the voice of the children with CLP may be reduced, but not in all cases. In our experiments there was only little difference in intonation and prosody between the CLP and the control children. In experiments of Haderlein [Hader 07a], the difference of the voice quality could be exploited to improve the intelligibility assessment of patients with severe voice disorder because the primary voice signal was disturbed. Hence, the extracted fundamental frequency features were significantly different to the ones of a control group, i.e., the prosodic feature set is more suitable for the assessment of voice disorders than the assessment of articulation disorders. Thirdly, prosody in children is related to the personality of the child, e.g. a shy child normally speaks with a low energy and rather monotone. This, however, is not clearly dependent on the articulation skills as seen by the clinicians during the examination of the children.

Since the WR as single feature proved to have the highest correlations with the mean opinion score of the experts, we evaluated this score on the CLP data and on the control groups in elementary school age. Table 6.11 displays the mean age and the mean WR. Note that all control groups (*Michael-Poeschke-06*, *Michael-Poeschke-07*, *Michael-Poeschke-08*, *Hannover-07*, *Karlsruhe-07*, and *Leipzig-07*) are age-matched

feature	prediction SVR		reference raters
	r	ρ	
word recognition rate	**0.89**	**0.92**	all raters
word recognition rate	0.86	0.89	rater M
+ maximum F0MinWord	0.83	0.89	rater M
+ mean PauseAfterWord	0.84	0.89	rater M
+ variance F0MeanWord	0.86	0.89	rater M
+ maximum EnergyMeanWord	**0.88**	**0.90**	rater M
word recognition rate	**0.80**	**0.84**	rater L
word recognition rate	0.79	0.85	rater S
+ maximum F0MinWord	**0.84**	**0.89**	rater S
word recognition rate	**0.85**	**0.82**	rater W

Table 6.10: Prediction of the experts' scores by different feature sets on the *CLP-Intel2* database

group	location	#	avg. age	WR
Michael-Poeschke-06	Erlangen	89	8.8 ± 1.3	60.5 ± 10.7
Michael-Poeschke-07	Erlangen	76	8.5 ± 1.4	62.7 ± 10.4
Michael-Poeschke-08	Erlangen	157	8.4 ± 1.2	61.1 ± 9.0
Hannover-07	Hannover	126	8.6 ± 1.1	63.7 ± 10.0
Karlsruhe-07	Karlsruhe	131	8.3 ± 1.1	64.1 ± 8.6
Leipzig-07	Leipzig	61	7.9 ± 2.1	59.0 ± 9.8
CLP-School	Erlangen	59	8.5 ± 1.6	52.2 ± 15.2

Table 6.11: Comparison of the distribution of the WR in the control groups from all over Germany and the patient group collected in Erlangen

age	controls		patients	
	#	WR	#	WR
six	61	54.2 ± 7.3	16	43.9 ± 10.0
seven	188	59.1 ± 9.0	9	42.4 ± 10.5
eight	155	63.2 ± 9.1	11	54.0 ± 18.5
nine	127	65.8 ± 8.5	9	57.7 ± 15.3
ten	94	68.0 ± 8.7	15	62.2 ± 12.4

Table 6.12: Overview on the WR in the control and the patient group

ID	cleft	age			WR		
	type	1st	2nd	diff.	1st	2nd	diff.
625	UCLP	15.9	16.6	0.7	65.7	69.7	4.0
4958	BCLP	12.7	13.9	1.2	64.7	70.7	6.1
5235	UCLP	12.8	14.0	1.2	71.7	76.8	5.1
12163	UCLP	9.3	10.1	0.8	34.3	28.3	-6.1
15017	UCLP	7,9	8.9	1.0	51.5	69.7	18.2
15552	UCLP	7.7	8.7	1.0	38.4	34.3	4.0
17661	CL	6.9	8.0	1.1	44.4	52.5	8.1
19646	UCLP	6.7	7.8	1.1	31.3	34.3	3.0
19793	UCLP	16.0	16.9	0.9	29.3	26.7	-3.0
21402	UCLP	6.0	6.9	0.9	45.5	48.5	3.0
24219	CP	6.2	6.7	0.5	45.5	35.4	-10.1
27159	UCLP	4.7	5.7	1.0	35.4	37.4	1.0
avg.	-	9.4	10.3	0.9	46.5	48.6	2.1
stddev.	-	3.9	3.9	0.2	14.2	18.6	7.4

Table 6.13: The children in the *CLP-Progression* dataset were recorded a second time after one year. Hence, their progress within this year can be compared.

and the patient group (*CLP-School*), i.e., there is no significant difference between each of the control groups to the patient group according to their age distribution. However, this is not true within all control groups. For example, *Leipzig-07* may not be compared to *Hannover-07* since their age distributions differ significantly. Mutual comparison between the control groups and the patient group revealed that the distribution in the control groups is always significantly higher ($p < 0.003$) than in the patient group. The overlap between patient and control groups is caused by the fact that children with CLP must not necessarily have a reduced intelligibility. Hence, children with normal WR also occur in the patient group.

In order to define age-dependent WR values, we put all control children together and formed a group for each age. So we created five groups for the ages six to ten for the control children and the CLP children. Table 6.12 lists the results. The difference between the means of the patient and control groups is about 10 percent points in all age groups. The variance of the WR in the patient children is also higher than in the control groups. In the control group, a significant improvement in WR is found each year ($p < 0.01$). This was expected since it is a known fact in the literature [Wilpo 96]. In the patient group, however, these significant improvements are missing. Between two subsequent age groups, no significant difference is found at all, even between seven and eight years ($p > 0.09$). This is of course related to the small number of children in the groups, but also to the fact that the speech disorder reduces the normal improvement in speech intelligibility, i.e., the speech of these children is not appropriate for their age. Furthermore, one has to consider that children stop coming to the follow-up care as soon as their speech is regarded as normal. Hence, only the children with speech disorder are recorded. The mended children do not appear in these statistics.

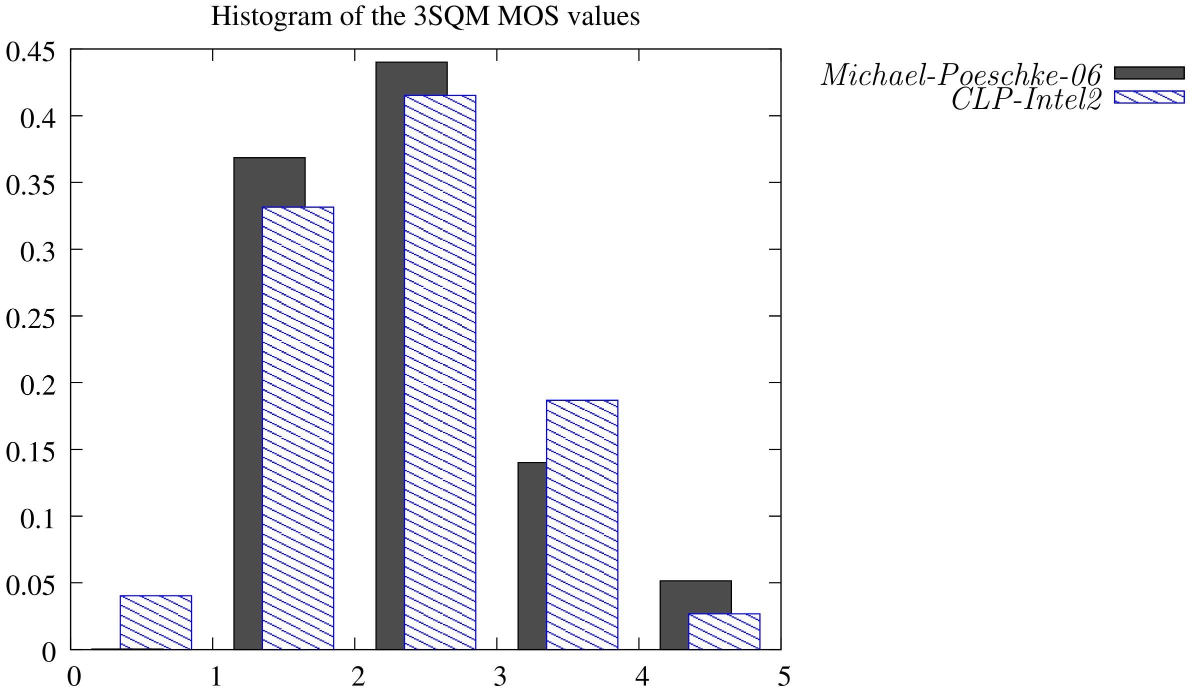


Figure 6.4: Mean Opinion Scores (MOS) obtained by the 3SQM [OPTI 04] procedure on the *CLP-Intel2* patient group and the *Michael-Poeschke-06* control group

Table 6.13 shows the children of the *CLP-Progression* group. The children were recorded twice with one year in between. About half of the children gain 3 to 5 percent points which is normal (cf. Table 6.12). However, some of the children (IDs 12163, 19793, 24219, and 27159) show little progress or even degradation. Perceptual evaluation confirmed this finding. In case of ID 24219, the decline in WR is amplified by a slight noise in the signal. Note that the children with no or just little progress are the children whose WR is already far below the age-matched average. Three of the children (IDs 4958, 15017, and 17661) improved above the age-matched average. Again, this finding could be perceptually confirmed. Especially, ID 15017 showed a lot of improvement. In average, the CLP children gain just 2.1 percent points which is lower than the children in the control group do. As in Table 6.12, the standard deviation is high, since some children improve and some don't. The group is too small to find significant differences within one year, but since all findings were perceptually confirmed, we believe that the method is suitable for documentation and screening of children's speech disorders.

A standard procedure to measure the quality of a speech channel is the Perceptual Evaluation of Speech Quality (PESQ) [Beere 02a, Beere 02b]. It is standardized by the International Telecommunication Union (ITU) [ITU 01]. The Result of the procedure is a score which was previously trained with the mean scores of a panel of experts. This Mean Opinion Score (MOS) is defined on a scale where 1.0 is worst and 5.0 best. Since the reference signal and the distorted signal are required to compute the PESQ MOS score, it cannot be applied to any of our datasets.

In contrary to PESQ, 3SQM can compute the MOS score without the need of the reference signal [OPTI 04]. Hence, it can be applied to the data presented here easily. 3SQM is also standardized by the ITU [ITU 04]. Details on the evaluation algorithm cannot be provided here. The implementation of 3SQM used in this work was provided binary by Opticom$^{\mathrm{TM}}$.

Figure 6.4 shows a histogram of the MOS, computed with the 3SQM algorithm for the *CLP-Intel2* and the *Michael-Poeschke-06* data sets. Mean and standard deviation are equal on both datasets (2.3 ± 0.9), i.e., both datasets are recorded in comparable audio quality. The result, however, also shows that the 3SQM algorithm is not applicable to determine the speech intelligibility in children with CLP.

6.3 Assessment of the Pronunciation

A major part of this work is the automatic evaluation of the pronunciation of children with CLP on frame, phoneme, word, and speaker level. Pronunciation evaluation is a typical classification problem since each frame, phoneme, or word has to be assigned either to class correct ("O") or to class wrong ("X"). This decision is performed by a classifier. Previously, two important classifiers were presented: The SVMs (cf. Chapter 4.2.2) and the GMMs (cf.Chapter 4.2.4). However, in order to be more flexible, this work employs more than these two classifiers. In the following, also classifiers of the WEKA toolbox [Witte 05] are used:

- **OneR:** The classifier divides the numeric features — also often called attributes in machine learning — into intervals which contain only observations — also called instances — of one class. In order to prevent over-fitting, mixed intervals are also allowed. However, each interval must hold at least a given number of instances in the training data. Then a decision rule for classification is created for each attribute. At the end of the training procedure, the attribute is selected for the classification which has the highest accuracy on the training set [Holte 93].

- **DecisionStump:** DecisionStumps are very simple classifiers. They are commonly used if training has to be performed often like in boosting algorithms. The classifier selects one attribute and a threshold or decision value to perform the classification. Selection is performed with correlation in the numeric case and entropy in the nominal case. Then, the selection value with the highest classification rate on the training set is determined.

- **LDA-Classifier:** The LDA-Classifier is also called "ClassificationViaRegression" [Frank 98a]. It basically determines a feature transformation matrix, like described in Chapter 4.2.3, and reduces the dimension to one. If no nominal class information but a numeric class is given, a Multiple Regression Analysis (cf. Chapter 4.2.1) is performed to reduce the dimension to one. Then, a simple threshold can be chosen to perform the classification. Again, the threshold is determined on the training set according to the best classification rate.

- **NaiveBayes:** The naïve Bayes classifier is trained according to Bayes' decision rule [Niema 03, p.315]. As probability density function a unimodal Gaussian mixture is chosen [John 95]. This classifier is equivalent to a GMM classifier with just one Gaussian distribution (cf. Chapter 4.2.4) with equal prior probabilities.

- **J48:** The J48 is an implementation of a C4.5 decision tree [Quinl 93]. In order to build a C4.5 decision tree, all instances in the data set are used to create a set of rules. Later on, the rules are pruned in order to reduce their number. Subsequently, a tree is generated which holds one simple decision rule concerning only one attribute, i.e., a DecisionStump in every node. At the leaves of the tree a class label is assigned. Classification is then performed starting from the tree root according to rules in the node. At the end of the classification, a leaf is reached which assigns the class to the observation.

- **PART:** In order to modify rules for a decision tree, two dominant approaches exist. The first one is dropping rules like the J48 tree does. The second one extends rules by replacing one or multiple rules by a better more refined rule. PART generates partial trees using both approaches and merges them later on. According to [Frank 98b] this method is much faster in training compared to J48 while having a similar or even better recognition accuracy.

- **RandomForest:** This kind of classifier is composed of multiple trees which are created randomly. For each tree a random subset of the training data is chosen. Then, a random subset of attributes is selected to be used in the tree. At each node, features are picked at random to determine the rule of the actual node. The rule which creates the best split for the current subset is computed. Such a random tree may not be pruned. The fusion of a random number of trees is then composed to a random forest [Breim 01].

- **AdaBoost:** Boosting is a common procedure to enhance simple classifiers. The idea of boosting is to join many weak classifiers to one single strong classifier. This is achieved by training in several iterations. In the following iteration, the data are re-weighted. Previously wrongly classified instances get a higher weight while correctly classified ones get a reduced weight. In this manner the classifiers adapt to the misclassified instances. Therefore, the algorithm is called AdaBoost [Freun 96].

In the following, the assessment on the completely transliterated dataset *CLP-Phone-Eval* is presented. Next, the methods developed for the *CLP-Phone-Eval* dataset are applied to the *CLP-Phone-Eval2* data where only the reference as pictures is known.

6.3.1 Semi-Automatic Assessment

The experiments on the pronunciation evaluation of the children with CLP on transliterated and non-transliterated data are similar. In fact, the experimental setup is the same. However, in the *CLP-Phone-Eval* data, some preprocessing steps have to be performed manually since no further knowledge about the test is available.

Experimental Setup

After the recording the data, they have to be preprocessed which is the first step in a classification system (cf. [Niema 03, p.26]). In the transliterated case, the spotting

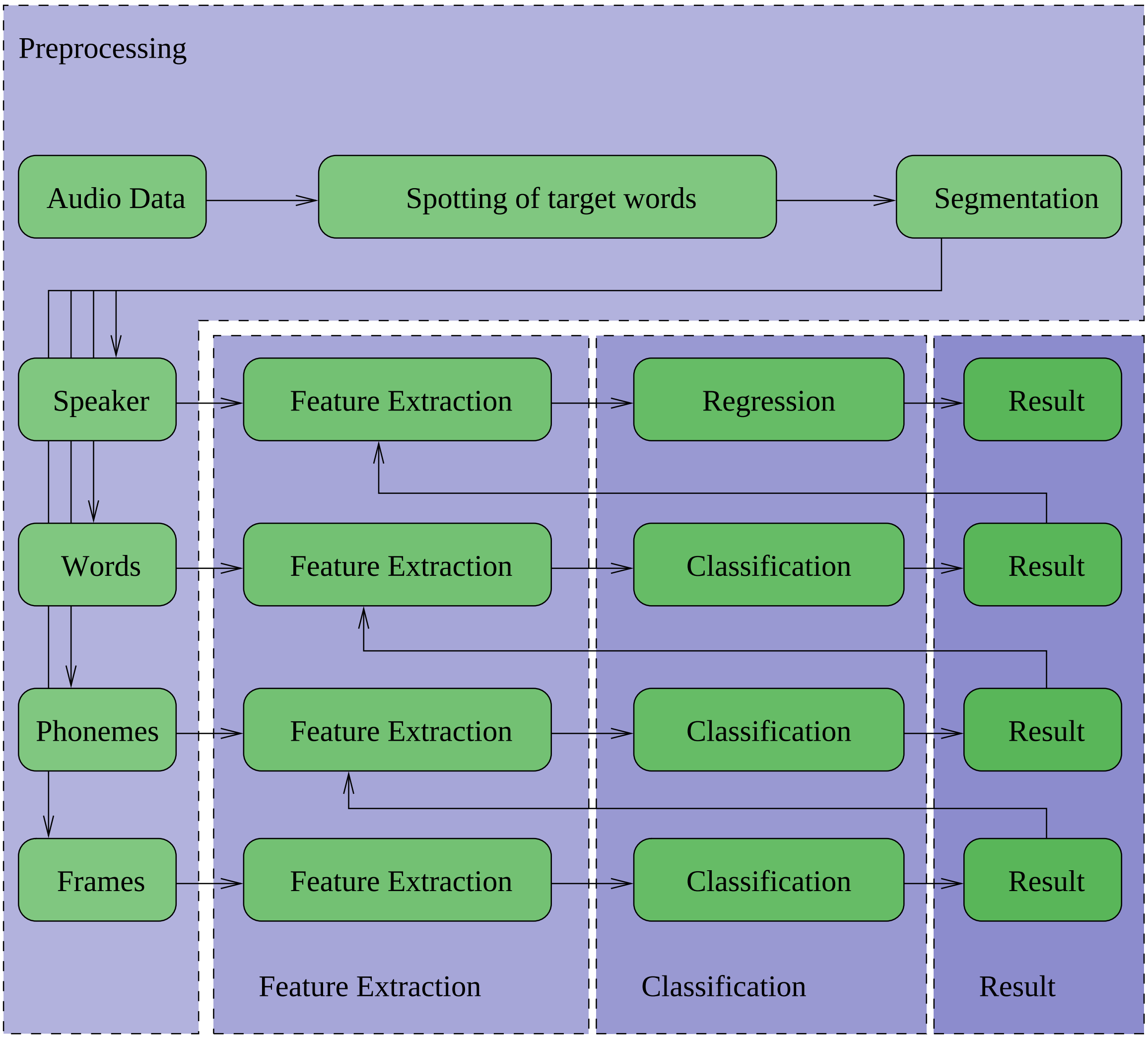


Figure 6.5: Experimental setup for the pronunciation assessment: After the preprocessing, the data are segmented to frame, phoneme, word, and speaker level. Results of the respective previous level are combined to the next level.

of target words was performed in the following manner: In a first segmentation step, the data was segmented automatically at long pauses. Next, all words which were spoken in the segments were transliterated, i.e., noted down by a naïve listener. During the transliteration procedure, all segments — which are also called "turns" in the literature — were cleaned from speech of the instructor of the test manually. The turns in which only the instructor was speaking were saved for preliminary experiments for speaker identification (cf. Chapter 4.2.4 and Chapter 6.3.2).

Next, the words which correspond to the PLAKSS words had to be identified. One of the main problems was that the pictograms on the slides of the PLAKSS test (cf. Chapter 5.1) cannot be projected one-to-one onto words. Most of the pictograms are ambiguous. Hence, a list of possible and common word alternatives has to be defined (cf. Appendix A.1.2). In order to get a preliminary mapping, the transliteration of each test of all the transliterated data (262 children from *CLP-02-06*, *Preschool*, and *Michael-Poeschke-06*) were force-aligned to the sequence of the

label	level	#	description
RecAcc	speaker	2	Accuracy of the speech recognition (WA and WR)
2-D Sammon Coordinates	speaker	2	Coordinates on a 2-D Sammon map
3-D Sammon Coordinates	speaker	3	Coordinates on a 3-D Sammon map
ProsFeat	word	37	Features based on the energy, the F_0, pauses, and duration to model the prosody of the speaker
PronFexW	word	7	Pronunciation features (PronFex) to score the correctness of the current word
PronFexP	phoneme	6	Features to score the correctness of the pronunciation (PronFex) of the current phone
TEP	phoneme	1	Teager Energy Profile to detect nasality in vowels
MFCCs	frame	24	Mel Frequency Cepstrum Coefficients plus first derivatives

Table 6.14: Overview on the feature sets used in this work (cf. Chapter 4.2.4)

PLAKSS words. Deletions and insertions were weighted with one while the substitutions were weighted by the Levenshtein difference between the two words divided by the length of the longer word. With this procedure we implicitly assumed that the children read the pictograms from left to right and that sensible word alternatives have a similar spelling. Both Assumptions were true in most cases. In order to ensure a correct correspondence between word alternatives and PLAKSS words, all mappings and their context were checked manually, i.e. $262 \times 99 = 25,938$ mappings were looked at and corrected by a human annotator in many hours of work.

With common word alternatives and fragments defined for every PLAKSS word, the word spotting was performed on the transliteration. Using forced alignment of the transliteration, the word and phoneme boundaries were determined (cf. Chapter 4.2.4). Then the data were segmented accordingly to word and phoneme level. Frame level data was created from the phoneme level data as described in Chapter 4.1.3. Figure 6.5 gives an overview of the procedure.

The subsequent steps of the classification system employed here are computed in parallel on four levels — frame, phoneme, word, and speaker level. Features are extracted on each of the levels. The different feature sets and their respective level are listed in Table 6.14. On frame level only *MFCCs* are used. However, after the training of a classifier, the training data is classified with the classifier, and functional features are computed. These functional features are then plugged into the next higher level of evaluation (here: phoneme level). As functionals the mean, the maximum, the minimum, the standard deviation, the sum, the count, the relative count, and the product of the classification result are computed. This kind of functional features can be computed for any kind of classifier and yield an abstract view of the lower level data to the higher-level classifier.

Next, the features of the respective level are fed to a classifier. Using the features of the training data, the classifier is first trained. The output of the classifier is either a class or a probability for each class. Classification is performed in favor of the class with the highest probability. While the classification on speaker, word, and phoneme level is performed by a single classifier, preliminary experiments quickly showed that phoneme-dependent classifiers are required on frame level. Since the correspondence between the data and the phonemes is clear due to the forced alignment, this procedure is valid without any restrictions even for the fully automatic assessment procedure as shown in the following.

Again, data were sparse which forced computation of all experiments in leave-one-speaker-out conditions. Hence, some of the features have to be recomputed in every iteration because they rely on statistics which are also estimated from the training data. *PronFexW* and *PronFexP* are calculated in each iteration newly since the training and test set must remain disjoint. The result of the evaluation is denoted as absolute recognition rate (RR) and class-wise averaged recognition rate (CL) since the distribution of the classes in the data is not balanced. The RR is computed as the total number of correctly recognized items divided by the number of total items in the test set. Hence, the rate is biased by the distribution of the classes in the data, i.e. if 99 % of the data belong to class “O” (here: correctly pronounced) and only 1 % to class “X” (wrongly pronounced), a classification rule which always classifies an observation to class “O” would always yield 99 % RR. To alleviate this problem, a second measure — the CL — is introduced. The CL is determined as the average of the recognition rate per class which is also often referred to as “recall”. This averaged recall would yield 50 % CL in the previously mentioned example. A rate of 50 % is considered as chance border in a two-class problem since a random decision rule converges against this number if sufficient experiments are performed. Again, the CL is also not perfect: If, for example, class “X” would have a recall of 100 % while class “O” just has a recall of 49 %, this would yield a CL of 74.5 % although only half of the data were correctly recognized, i.e. RR is 50 %. Hence, evaluation results should always contain both numbers. The use of an F-measure is set aside in this work since there are multiple conflicting definitions in the literature [Rijsb 79, Yang 99, Schul 07].

In the following the experiments on hypernasality (HN), nasalized consonants (NC), laryngeal replacement (LR), pharyngeal backing (PB), and weakened plosives (WP) are presented. Experiments on interdentalization and lateralization were not performed since both articulation errors are neither connected to CLP nor to the speech intelligibility.

Hypernasalization

Hypernasalization experiments were conducted with all available feature sets. On frame level, GMM classification was performed on all phonemes simultaneously while the other classifiers were trained for each phoneme individually. For individual phoneme training, the other classifiers have much fewer training data than the GMMs. While the RR grows steadily with an increasing number of densities in the GMM

classifier	CL	RR
OneR	50.0 %	**99.0 %**
DecisionStump	54.6 %	87.0 %
LDA-Classifier	55.6 %	91.8 %
NaiveBayes	54.0 %	94.4 %
J48	53.8 %	90.5 %
PART	55.0 %	90.3 %
RandomForest	49.8 %	98.7 %
SVM	**56.8 %**	91.4 %
GMM (2 densities)	52.2 %	81.7 %
GMM (5 densities)	47.6 %	79.1 %
GMM (10 densities)	48.3 %	84.0 %
GMM (15 densities)	47.2 %	84.4 %

Table 6.15: Overview on the CL and RR of hypernasalization obtained by different classifiers on the *CLP-Phone-Eval* dataset on frame level

classifier, the CL drops. With more than two densities, the CL already drops below chance level. The CL with 15 densities is worst. Recognition with a classifier for each phoneme works much better. The recognition rates are significantly better than chance except for the OneR classifiers and the RandomForests. However, both classifiers have very high RRs (99.0 % and 98.7 %).

Table 6.16 lists the classification rates on phoneme level for the *CLP-Phone-Eval*database. The application of just the *TEP* to the data yields much lower recognition rates than found in the literature. Of course, the features were not designed for children's speech, but the recognition here is much worse compared to the numbers in the literature. However, one has to keep in mind, that this work is the first one which does the complete segmentation semi-automatically and not manually. Moreover, this is the first time that the methods are applied to real pathologic data of children. Hence, a reduction of the recognition rates was expected. The OneR classifier and the RandomForest lie in the same range, as described in [Cairn 96a]. Their CL, however, is at chance level. In fact, both classifiers assigned virtually always label "O". Unfortunately, Cairns does not provide exact information about the dataset he used, e.g. the distribution of this classes. In real data, as presented here, the non-hypernasal case is predominant. In the *CLP-Phone-Eval* data, about 50 of the 1916 words of the dataset contain hypernasal vowels (cf. Table A.7). The actual training data is in the range of seconds. With the other classifiers, recognition rates are much lower. The best CL is found for the LDA-Classifier with 59.2 %.

MFCCs perform significantly worse than the *TEP* features in CL. The best CL is found with the NaiveBayes classifier based on functionals of the MFCC features computed from the output of a frame level MFCC NaiveBayes Classifier. RR is highest with 99 % in the RandomForest and OneR classifiers. Their CL, however, is again at chance level. The best CL is found obtained with the NaiveBayes classifier (56.9 %). Combination of both feature sets is beneficial: The CL can be improved to up to 62.9 % with the DecisionStump classifier. Additional use of pronunciation features improves only the RR in most classifiers. The CL is only improved in the PART and

classifier	feature sets	CL	RR
OneR	*TEP*	50.0 %	**97.7 %**
DecisionStump	*TEP*	53.9 %	16.6 %
LDA-Classifier	*TEP*	**59.2 %**	56.4 %
NaiveBayes	*TEP*	52.7 %	45.3 %
J48	*TEP*	55.7 %	50.9 %
PART	*TEP*	54.7 %	50.9 %
RandomForest	*TEP*	48.1 %	94.2 %
SVM	*TEP*	49.4 %	73.9 %
OneR	OneR (*MFCCs*)	50.0 %	**99.0 %**
DecisionStump	DecisionStump (*MFCCs*)	53.9 %	85.9 %
LDA-Classifier	LDA-Classifier (*MFCCs*)	52.9 %	92.4 %
NaiveBayes	NaiveBayes (*MFCCs*)	**56.9 %**	87.5 %
J48	J48 (*MFCCs*)	53.6 %	91.7 %
PART	PART (*MFCCs*)	52.2 %	90.9 %
RandomForest	RandomForest (*MFCCs*)	50.0 %	**99.0 %**
SVM	SVM (*MFCCs*)	56.2 %	90.6 %
OneR	OneR (*MFCCs*), *TEP*	49.9 %	93.4 %
DecisionStump	DecisionStump (*MFCCs*), *TEP*	**62.9 %**	87.1 %
NaiveBayes	NaiveBayes (*MFCCs*), *TEP*	51.2 %	82.6 %
J48	J48 (*MFCCs*), *TEP*	53.9 %	93.6 %
PART	PART (*MFCCs*), *TEP*	51.5 %	92.0 %
SVM	SVM (*MFCCs*), *TEP*	53.2 %	91.1 %
OneR	OneR (*MFCCs*), *TEP*, *PronFexP*	49.8 %	**98.7 %**
DecisionStump	DecisionStump (*MFCCs*), *TEP*, *PronFexP*	**60.6 %**	82.8 %
NaiveBayes	NaiveBayes (*MFCCs*), *TEP*, *PronFexP*	51.2 %	94.5 %
J48	J48 (*MFCCs*), *TEP*, *PronFexP*	53.9 %	93.6 %
PART	PART (*MFCCs*), *TEP*, *PronFexP*	55.1 %	93.1 %
SVM	SVM (*MFCCs*), *TEP*, *PronFexP*	54.9 %	91.6 %

Table 6.16: Recognition rates on phoneme level for hypernasalization on the *CLP-Phone-Eval* dataset: The terms in brackets describe the input data to the respective classifier. The best result for each combination of feature sets is printed in bold face.

the SVM. The best CL with all phoneme level feature sets has the DecisionStump with 60.6 %.

In Table 6.17 the results on word level are displayed. Again, the RandomForest and the OneR classifier show the highest RRs. In contrary to the OneR, the RandomForest is above chance level with 51.2 % CL and 96.9 % RR. The best result using just *MFCCs* is obtained with the SVM with 52.5 % CL. Additional use of the *TEP* features increases the CL with the DecisionStump to 57.7 % CL while the RR stays in the same range as with the SVM. Significant improvement is achieved by the addition of *PronFexP* features on phoneme level and feature selection with Maximum R ($p < 0.05$). This yields a recognition of 60.6 % CL. Additional use of word level features like *PronFexW* and *ProsFeat* did not bring any further improvements.

On speaker level, significant correlations between the percentage of the detected hypernasalized vowels and the annotated number are found as presented in Table 6.18. In the opinion of the author, the correlation between the percentages of annotated and detected words is more reliable than their number, because each child uttered a different number of target words which could be mapped onto the words of the PLAKSS test. Due to the prior processing, the number of detected target words varies between the different speakers. If just the plain numbers would be correlated, a good correlation would mean that the classifier detected a matching number of words. The number of detected words, however, is dependent on the total number of words. The fewer words, the fewer is the chance to detect words as wrongly pronounced. This fact might improve correlation artificially. Therefore, the percentage of detected and annotated words were chosen to compute the correlation.

On the one hand, the RandomForest has a high correlation because it had a very high RR on word level. On the other hand, the DecisionStump also has a significant correlation although its RR was significantly lower ($p < 0.001$). The CL of the DecisionStump, however, was significantly higher than the CL of the RandomForest. Hence, it can be concluded that both measures — the CL and the RR — are of importance to find a good predictor of hypernasality. Insignificant correlations are not reported in Table 6.18. Addition of further information is beneficial for the correlation. The *RecAcc* features improve the correlation to 0.78. If the coordinates of the Sammon map are further supplied to the prediction, the correlation further increases to 0.89 which is also highly significant ($p < 0.001$).

Nasalized Consonants

Table 6.19 lists the recognition performance of the different classifiers on frame level on the *CLP-Phone-Eval* dataset. The best RR is obtained with the OneR classifier which basically assigns all vectors the label “O”. Hence, RR is very high. The CL, however, is very close to chance level. With the NaiveBayes classifier, i.e. a unimodal GMM, the best CL with 62.0 % was achieved. The RR of the NaiveBayes classifier is second-best right behind the OneR classifier. This is significantly better than the CL of the DecisionStump ($p < 0.005$) and the RR of the LDA-Classifier ($p < 0.05$). On average the NaiveBayes classifier performs best. Classification for individual phonemes proved to be significantly better than for all phonemes together: The GMM classifier trained for all phonemes shows significantly worse CL and RR ($p < 0.001$).

classifier	feature sets	CL	RR
OneR	OneR (OneR (*MFCCs*))	49.9 %	**96.8 %**
DecisionStump	DecisionStump (DecisionStump (*MFCCs*))	45.0 %	68.8 %
LDA-Classifier	LDA-Classifier (LDA-Classifier (*MFCCs*))	51.7 %	84.2 %
NaiveBayes	NaiveBayes (NaiveBayes (*MFCCs*))	46.3 %	61.8 %
J48	J48 (J48 (*MFCCs*))	50.2 %	71.0 %
PART	PART (PART (*MFCCs*))	46.1 %	65.8 %
RandomForest	RandomForest (RandomForest (*MFCCs*))	51.4 %	**96.9 %**
SVM	SVM (SVM (*MFCCs*))	**52.3 %**	70.2 %
OneR	OneR (OneR (*MFCCs*), *TEP*)	50.8 %	**95.8 %**
DecisionStump	DecisionStump (DecisionStump (*MFCCs*), *TEP*)	**57.7 %**	69.5 %
NaiveBayes	NaiveBayes (NaiveBayes (*MFCCs*), *TEP*)	43.4 %	54.8 %
J48	J48 (J48 (*MFCCs*), *TEP*)	51.3 %	76.2 %
PART	PART (PART (*MFCCs*), *TEP*)	45.2 %	70.0 %
SVM	SVM (SVM (*MFCCs*), *TEP*)	52.7 %	71.0 %
DecisionStump	DecisionStump (DecisionStump (*MFCCs*), *TEP*, *PronFexP*)	**56.8 %**	62.0 %
SVM	Maximum R (SVM (Maximum R (SVM(*MFCCs*), *TEP*, *PronFexP*)))	**60.6 %**	69.4 %
SVM	Maximum R (PART (Maximum R (PART(*MFCCs*), *TEP*, *PronFexP*)), *PronFexW*, *ProsFeat*)	54.1 %	78.3 %

Table 6.17: Recognition rates on word level for hypernasalization on the *CLP-Phone-Eval* dataset

feature sets	r
RandomForest (RandomForest (RandomForest(*MFCCs*)))	0.75 (**)
DecisionStump (DecisionStump (DecisionStump(*MFCCs*), *TEP*))	0.64 (**)
DecisionStump (Maximum R (DecisionStump (Maximum R (DecisionStump(*MFCCs*), *TEP*, *PronFexP*)))	0.57 (**)
RandomForest (RandomForest (RandomForest (*MFCCs*))), *RecAcc*	0.78 (**)
RandomForest (RandomForest (RandomForest (*MFCCs*))), *RecAcc*, *2-D Sammon Coordinates*	0.87 (**)
RandomForest (RandomForest (RandomForest (*MFCCs*))), *RecAcc*, *3-D Sammon Coordinates*	**0.89** (**)

Table 6.18: Correlations on speaker level between the percentage of detected words and the annotated percentage for hypernasalization on the *CLP-Phone-Eval* dataset. (**) marks significant correlations with $p < 0.01$.

classifier	CL	RR
OneR	50.8 %	**94.2 %**
DecisionStump	59.8 %	62.7 %
LDA-Classifier	59.0 %	78.7 %
NaiveBayes	**62.0 %**	80.2 %
J48	56.1 %	74.1 %
PART	57.9 %	75.3 %
RandomForest	53.2 %	92.3 %
SVM	57.1 %	76.4 %
GMM (2 densities)	56.2 %	53.4 %
GMM (5 densities)	53.4 %	63.2 %
GMM (10 densities)	52.9 %	65.1 %
GMM (15 densities)	49.9 %	73.3 %

Table 6.19: Overview on the CL and RR of nasalized consonants obtained by different classifiers on the *CLP-Phone-Eval* dataset on frame level

Note that the *TEP* features were not evaluated since they are only defined on voiced speech, i.e. vowels.

Application of *MFCCs* on phoneme level gives CL rates of up to 66.7 %. RR, however, is only 53.9 %. OneR and RandomForest yield high RRs, but their CL is again very close to chance. The *PronFexP* features also perform quite well. Using an J48 classifier, a CL rate of 67.5 % is achieved. However, RR is only 56.6 %. Combination of *MFCCs* with the *PronFexP* features yields recognition rates which are comparable to using only one of the feature sets. While the *PronFexP* features have slightly better discrimination between the classes, i.e., a higher CL, the RR is better in the *MFCCs*. Both effects can be brought to a single classifier by application of further improvement techniques: Feature selection with Maximum R (cf. Chapter 4.2.3) and additional boosting of the J48 classifier gives a classifier which is close to the DecisionStump in CL with 67.9 % while having a significantly higher RR with 79.8 % ($p < 0.001$).

Recognition on word level is more difficult than on phoneme level. On the one hand, there are many possible combinations to choose classifiers on frame, phoneme, and word level. Moreover, application of the same classifier on each level yields only suboptimal results, as displayed in Table 6.21. The table lists only the best combinations which were achieved. The best RRs which could be obtained are around 80 %. The best CL using *MFCCs* was only 63.6 %. The investigation of only word level features yielded just weak classification: The best CL with the *ProsFeat* features was 52.3 %. Using *PronFexW* features, the best CL was 53.9 %. Combination of all features gave an improvement in CL to 58.4 % which is still lower than the recognition with just *MFCCs*.

Table 6.22 shows the results on speaker level. Using just MFCC features, a correlation of 0.71 between the percentage of words which contain nasalized phonemes and the annotated words is found. Further improvement in the regression is achieved if the *RecAcc* features are added to the regression. Addition of *2-D Sammon Coordinates* or *3-D Sammon Coordinates* enhances the regression even more. The application of all features on word level with an SVM gave a correlation of 0.47. If all available information is provided to the prediction, a correlation of 0.85 is achieved.

Laryngeal Replacement

The multi-level classification system as presented in Figure 6.5 can also be employed to detect laryngeal replacements. Table 6.23 lists results on the *CLP-Phone-Eval* data on frame level. Again very high RRs are found with the OneR classifier and the RandomForest. The corresponding CLs are, again, at chance level. The best CL with *MFCCs* is found with the PART with a CL of 59.8 % and a RR of 91.3 %. Also for the laryngeal replacement, the evaluation with individual classifiers for each phoneme is beneficial. The CL is significantly higher in the PART classifier compared to the GMMs trained with all observed phonemes ($p < 0.001$).

On phoneme level, *MFCCs* yield good classification results (cf. Table 6.24). With the J48 a classification with 61.7 % CL is achieved. Even better is classification using an SVM on phoneme level and a PART classifier on frame level with a CL of 69.5 %. Additional use of pronunciation features could improve the RR, but not the CL in

classifier	feature sets	CL	RR
OneR	OneR (*MFCCs*)	51.9 %	**94.6 %**
DecisionStump	DecisionStump (*MFCCs*)	**66.7 %**	53.9 %
LDA-Classifier	LDA-Classifier (*MFCCs*)	60.2 %	72.5 %
NaiveBayes	NaiveBayes (*MFCCs*)	63.5 %	56.6 %
J48	J48 (*MFCCs*)	62.2 %	78.3 %
PART	PART (*MFCCs*)	62.6 %	77.2 %
RandomForest	RandomForest (*MFCCs*)	50.2 %	**95.6 %**
SVM	SVM (*MFCCs*)	56.1 %	75.5 %
OneR	*PronFexP*	49.9 %	**94.5 %**
DecisionStump	*PronFexP*	64.4 %	56.6 %
LDA-Classifier	*PronFexP*	59.3 %	60.2 %
NaiveBayes	*PronFexP*	53.7 %	26.6 %
J48	*PronFexP*	**67.5 %**	65.3 %
PART	*PronFexP*	61.3 %	61.5 %
RandomForest	*PronFexP*	52.7 %	**91.5 %**
SVM	*PronFexP*	61.0 %	63.5 %
OneR	OneR (*MFCCs*), *PronFexP*	52.8 %	94.0 %
DecisionStump	DecisionStump (*MFCCs*), *PronFexP*	**68.5 %**	52.2 %
LDA-Classifier	LDA-Classifier (*MFCCs*), *PronFexP*	63.9 %	74.2 %
NaiveBayes	NaiveBayes (*MFCCs*), *PronFexP*	64.6 %	70.0 %
J48	J48 (*MFCCs*), *PronFexP*	59.3 %	79.8 %
PART	PART (*MFCCs*), *PronFexP*	63.1 %	78.4 %
RandomForest	RandomForest (*MFCCs*), *PronFexP*	50.0 %	**95.1 %**
SVM	SVM (*MFCCs*),*PronFexP*	56.5 %	72.8 %
AdaBoost (J48)	Maximum R (NaiveBayes (*MFCCs*), *PronFexP*)	**67.9 %**	79.8 %

Table 6.20: Recognition rates on phoneme level for nasalized consonants on the *CLP-Phone-Eval* dataset

classifier	feature sets	CL	RR
OneR	OneR (OneR (*MFCCs*))	52.0 %	**79.7 %**
RandomForest	RandomForest (RandomForest (*MFCCs*))	51.4 %	**82.5 %**
DecisionStump	J48 (NaiveBayes (*MFCCs*))	**63.6 %**	63.6 %
OneR	*ProsFeat*	50.1 %	**81.5 %**
LDA-Classifier	*ProsFeat*	52.3 %	54.9 %
OneR	*PronFexW*	50.1 %	**81.4 %**
J48	*PronFexW*	53.9 %	68.3 %
SVM	J48 (NaiveBayes (*MFCCs*), *PronFexP*), *PronFexW*, *ProsFeat*	58.4 %	62.9 %

Table 6.21: Recognition rates on word level for nasalized consonants on the *CLP-Phone-Eval* dataset

feature sets	R
DecisionStump (J48 (NaiveBayes(*MFCCs*)))	0.70 (**)
DecisionStump (J48 (Maximum R (NaiveBayes (*MFCCs*))))	0.71 (**)
DecisionStump (J48 (Maximum R (NaiveBayes (*MFCCs*)))), *RecAcc*	0.82 (**)
DecisionStump (J48 (Maximum R (NaiveBayes (*MFCCs*)))), *RecAcc*, *2-D Sammon Coordinates*	0.83 (**)
DecisionStump (J48 (Maximum R (NaiveBayes (*MFCCs*)))), *RecAcc*, *3-D Sammon Coordinates*	0.84 (**)
SVM (J48 (NaiveBayes (*MFCCs*), *PronFexP*), *PronFexW*, *ProsFeat*)	0.47 (*)
SVM (J48 (NaiveBayes (*MFCCs*), *PronFexP*), *PronFexW*, *ProsFeat*), DecisionStump (J48 (Maximum R (NaiveBayes (*MFCCs*)))), *RecAcc*, *3-D Sammon Coordinates*	**0.85** (**)

Table 6.22: Correlations on speaker level between the percentage of detected words and the annotated percentage for nasalized consonants on the *CLP-Phone-Eval* dataset. (*) marks significant correlations with $p < 0.05$ and (**) correlations with $p < 0.01$.

classifier	CL	RR
OneR	50.0 %	**99.3 %**
DecisionStump	52.8 %	89.7 %
NaiveBayes	51.7 %	96.4 %
J48	59.7 %	91.2 %
PART	**59.8 %**	91.3 %
RandomForest	50.0 %	**99.6 %**
SVM	53.0 %	93.7 %
GMM (2 densities)	49.7 %	92.4 %
GMM (5 densities)	49.5 %	95.9 %
GMM (10 densities)	50.1 %	96.5 %
GMM (15 densities)	50.5 %	91.8 %

Table 6.23: Overview on the CL and RR of laryngeal replacement obtained by different classifiers on the *CLP-Phone-Eval* dataset on frame level

classifier	feature sets	CL	RR
OneR	OneR(*MFCCs*)	50.0 %	**99.6 %**
NaiveBayes	NaiveBayes(*MFCCs*)	52.8 %	88.7 %
J48	J48(*MFCCs*)	**61.7 %**	89.7 %
PART	PART(*MFCCs*)	59.8 %	91.3 %
RandomForest	RandomForest(*MFCCs*)	50.0 %	**99.6 %**
SVM	SVM(*MFCCs*)	58.6 %	90.0 %
SVM	PART(*MFCCs*)	**69.5 %**	88.9 %
SVM	PART(*MFCCs*), *PronFexP*	65.3 %	**92.6 %**

Table 6.24: Recognition rates on phoneme level for laryngeal replacement on the *CLP-Phone-Eval* dataset

classifier	feature sets	CL	RR
OneR	OneR (OneR (*MFCCs*))	50.0 %	**98.1 %**
RandomForest	RandomForest (RandomForest (*MFCCs*))	50.0 %	**98.1 %**
LDA-Classifier	LDA-Classifier (LDA-Classifier (*MFCCs*))	**59.1 %**	83.6 %
SVM	SVM (PART (*MFCCs*))	**62.0 %**	63.9 %
J48	LDA-Classifier (PART (*MFCCs*))	**63.8 %**	80.8 %
PART	LDA-Classifier (PART (*MFCCs*))	**63.8 %**	80.8 %
SVM	SVM (PART (*MFCCs*), *PronFexP*)	57.7 %	72.6 %
PART	SVM (PART (*MFCCs*), *PronFexP*)	60.0 %	81.1 %
SVM	SVM (PART (*MFCCs*), *PronFexP*), *PronFexW*	57.7 %	72.6 %

Table 6.25: Recognition rates on word level for laryngeal replacement on the *CLP-Phone-Eval* dataset

most classifiers. The best classifier representing this group was the SVM with 65.3 % CL.

Table 6.25 lists the results on word level on the *CLP-Phone-Eval* data. The OneR and the RandomForest classifiers yield the highest RR with 98.1 % RR. The CL of 50.0 % reveals that both classifiers assigned label “O” in all cases. The class “X” was never classified. With other classifiers, higher CLs are found. A combination of three LDA-Classifiers, for example, has a CL of 59.1 %. The best combination in terms of CL using *MFCCs* only is achieved with a PART classifier on frame level, an SVM on phoneme level, and either a J48 or a PART on word level with a CL of 63.8 %. Use of additional features on phoneme or word level did not bring any further advances. Therefore, the table presents only a small number of these combinations.

feature sets	R
J48 (LDA (PART (*MFCCs*)))	0.56 (*)
PART (LDA (PART (*MFCCs*)))	0.56 (*)
PART (SVM (PART (*MFCCs*), *PronFexP*))	0.50 (*)
PART (LDA (PART (*MFCCs*))), *RecAcc*	0.73 (**)
PART (LDA (PART (*MFCCs*))), *RecAcc*, *2-D Sammon Coordinates*	0.80 (**)
PART (LDA (PART (*MFCCs*))), *RecAcc*, *3-D Sammon Coordinates*	0.81 (*)

Table 6.26: Correlations on speaker level between the percentage of detected words and the annotated percentage for laryngeal replacement on the *CLP-Phone-Eval* dataset. (*) marks significant correlations with $p < 0.05$ and (**) correlations with $p < 0.01$.

classifier	CL	RR
OneR	50.1 %	**99.1 %**
DecisionStump	58.7 %	88.7 %
NaiveBayes	55.0 %	95.6 %
J48	**63.6 %**	91.4 %
PART	**63.2 %**	91.5 %
RandomForest	51.2 %	**98.6 %**
SVM	**66.0 %**	93.6 %
GMM (2 densities)	49.4 %	88.2 %
GMM (5 densities)	49.5 %	93.5 %
GMM (10 densities)	49.9 %	97.2 %
GMM (15 densities)	49.9 %	97.7 %

Table 6.27: Overview on the CL and RR of pharyngeal backing obtained by different classifiers on the *CLP-Phone-Eval* dataset on frame level

classifier	feature sets	CL	RR
OneR	OneR(*MFCCs*)	50.0 %	**99.6 %**
NaiveBayes	NaiveBayes(*MFCCs*)	65.2 %	92.9 %
J48	J48(*MFCCs*)	63.6 %	90.9 %
PART	PART(*MFCCs*)	**66.7 %**	95.6 %
RandomForest	RandomForest(*MFCCs*)	50.0 %	**99.6 %**
SVM	SVM(*MFCCs*)	63.7 %	89.8 %
OneR	OneR(*MFCCs*), *PronFexP*	51.9 %	**99.5 %**
SVM	SVM(*MFCCs*), *PronFexP*	**76.9 %**	88.6 %

Table 6.28: Recognition rates on phoneme level for pharyngeal backing on the *CLP-Phone-Eval* dataset

The percentage of detected words of the word classifiers showed significant correlations with the perceptively annotated words (cf. Table 6.26). Using only *MFCCs* yields a correlation of 0.56 with a combination of PART and LDA classifiers. Improvement is achieved by addition of the *RecAcc* features. Inclusion of further information from *2-D Sammon Coordinates* or *3-D Sammon Coordinates* results in a correlation of up to 0.81.

Pharyngeal Backing

For pharyngeal backing quite high recognition rates were achieved on frame level (cf. Table 6.27). With the SVM, a CL of 66.0 % and a RR of 93.6 % were found. Furthermore, the evaluation for individual phonemes outperforms the recognition with a single classifier for all phonemes. The GMM classifier shows much lower CLs than most of the individually trained classifiers.

Table 6.28 lists recognition rates on phoneme level. The best classifier in terms of CL with *MFCCs* was the PART. Additional improvement by combination with *PronFexP* features could be obtained with the OneR classifier and the SVM. The SVM yields a significant improvement in CL up to 76.9 % ($p > 0.001$). Further

classifier	feature sets	CL	RR
OneR	OneR (OneR (*MFCCs*))	50.0 %	**98.2 %**
RandomForest	RandomForest (RandomForest (*MFCCs*))	50.0 %	**98.2 %**
PART	PART (PART (*MFCCs*))	**59.1 %**	82.7 %
DecisionStump	DecisionStump (DecisionStump (*MFCCs*))	**59.8 %**	61.8 %
PART	PART (PART (*MFCCs*), *PronFexP*)	**59.6 %**	84.8 %
SVM	SVM (SVM (*MFCCs*), *PronFexP*)	**67.9 %**	61.8 %
SVM	SVM (SVM (*MFCCs*), *PronFexP*), *PronFexW*	**67.2 %**	60.2 %

Table 6.29: Recognition rates on word level for pharyngeal backing on the *CLP-Phone-Eval* dataset

feature sets	R
PART (PART (PART (*MFCCs*)))	0.46 (*)
PART (PART (PART (*MFCCs*))), *RecAcc*, *2-D Sammon Coordinates*	0.68 (*)
PART (PART (PART (*MFCCs*))), *RecAcc*, *3-D Sammon Coordinates*	**0.70** (*)

Table 6.30: Correlations on speaker level between the percentage of detected words and the annotated percentage for pharyngeal backing on the *CLP-Phone-Eval* dataset. (*) marks significant correlations with $p < 0.05$ and (**) correlations with $p < 0.01$.

combinations of classifiers and features did not yield any improvement and are hence not reported.

The recognition of pharyngeal backing works also well on word level (cf. Table 6.29). Application of *MFCCs* only resulted in a recognition of 59.1 % CL. Additional use of pronunciation features on phoneme level brings a further improvement to 67.9 % CL. Pronunciation and prosodic features on word level do not increase the recognition additionally.

The correlation between the detected and the annotated pharyngeal backing on the *CLP-Phone-Eval* data is 0.46. The correlation is further increased by combination with the *RecAcc* features and the *2-D Sammon Coordinates* to 0.68. If the *3-D Sammon Coordinates* are used, a correlation of 0.70 is achieved.

Weakened Plosives

Weakened plosives were detected well already on frame level (cf. Table 6.31). Note that weakening of the plosives in the region of Erlangen might also be related to the dialect and not to the clefting. Therefore, the amount of training data was higher compared to the other articulation disorders. The classification with the GMM

classifier	CL	RR
OneR	50.1 %	**97.8 %**
DecisionStump	**71.1 %**	81.8 %
NaiveBayes	62.1 %	90.8 %
J48	**66.8 %**	88.9 %
PART	**67.0 %**	87.6 %
RandomForest	50.9 %	**97.3 %**
SVM	**67.7 %**	88.2 %
GMM (2 densities)	56.1 %	75.4 %
GMM (5 densities)	53.9 %	59.6 %
GMM (10 densities)	53.3 %	76.2 %
GMM (15 densities)	52.4 %	74.7 %

Table 6.31: Overview on the CL and RR of weakened plosives obtained by different classifiers on the *CLP-Phone-Eval* dataset on frame level

classifier for all phonemes with a single classifier showed the best results compared to the other GMM classifiers of the other criteria. However, individual classification of all phonemes yielded much better CLs and RRs. The best CL was obtained with the DecisionStump with 71.1 %. The other classifiers, like J48, PART, and SVM, also have very high CLs.

In Table 6.32 the results for the recognition of weakened plosives on phoneme level are presented. The best CL using *MFCCs* only is 71.1 % with a DecisionStump classifier. Slightly lower recognition rates are achieved with the SVM with 67.7 % CL and the PART with 67.0 % CL. The additional use of *PronFexP* improves the recognition for most classifiers. The best CL, with a combination of *MFCCs* and *PronFexP* features, is 71.0 %.

Table 6.33 displays the results for weakened plosives on word level. The best recognition with just *MFCCs* is 66.1 % using a SVM. Further improvement is achieved by addition of *PronFexP*. The CL rises to 67.7 %. Prosodic information and more pronunciation features refine the recognition even more. A CL of 75.8 % is obtained.

The assessment of the weakened plosives works also well on speaker level. Using just the word level features, a correlation of 0.61 is achieved. Addition of the *RecAcc* features yields 0.73 correlation. The *3-D Sammon Coordinates* improve the correlation further to 0.81. The best correlation with 0.82 is created by application of two word level recognizers and the speaker level features.

6.3.2 Fully Automatic Assessment

In order to show that the segmentation of the audio data is also possible automatically, we investigated the speech data of the *CLP-Phone-Eval2* database. The experimental setup is very similar to Figure 6.5. However, the segmentation is performed automatically.

classifier	feature sets	CL	RR
OneR	OneR (*MFCCs*)	50.1 %	**97.8 %**
DecisionStump	DecisionStump (*MFCCs*)	**71.1 %**	81.8 %
LDA-Classifier	LDA-Classifier (*MFCCs*)	64.2 %	89.2 %
NaiveBayes	NaiveBayes (*MFCCs*)	62.1 %	90.8 %
J48	J48 (*MFCCs*)	66.8 %	88.9 %
PART	PART (*MFCCs*)	**67.0 %**	87.6 %
RandomForest	RandomForest (*MFCCs*)	50.9 %	**97.3 %**
SVM	SVM (*MFCCs*)	**67.7 %**	88.2 %
DecisionStump	SVM (*MFCCs*), *PronFexP*	**70.4 %**	88.5 %
LDA-Classifier	SVM (*MFCCs*), *PronFexP*	69.6 %	88.0 %
J48	SVM (*MFCCs*), *PronFexP*	**70.4 %**	88.5 %
PART	SVM (*MFCCs*), *PronFexP*	**70.4 %**	88.5 %
SVM	SVM (*MFCCs*), *PronFexP*	**71.0 %**	85.0 %

Table 6.32: Recognition rates on phoneme level for weakened plosives on the *CLP-Phone-Eval* dataset

classifier	feature sets	CL	RR
OneR	OneR (OneR (*MFCCs*))	50.1 %	**97.8 %**
RandomForest	RandomForest (RandomForest (*MFCCs*))	51.8 %	**94.2 %**
PART	PART (PART (*MFCCs*))	**63.0 %**	61.8 %
SVM	SVM (SVM (*MFCCs*))	**66.1 %**	62.3 %
LDA-Classifier	LDA-Classifier (SVM (*MFCCs*), *PronFexP*)	**67.7 %**	70.7 %
LDA-Classifier	Maximum R (LDA-Classifier (Maximum R (SVM (*MFCCs*), *PronFexP*)), *PronFexW*, *ProsFeat*)	**68.1 %**	76.9 %
SVM	Maximum R (LDA-Classifier (Maximum R (SVM (*MFCCs*), *PronFexP*)), *PronFexW*, *ProsFeat*)	**75.8 %**	68.1 %

Table 6.33: Recognition rates on word level for weakened plosives on the *CLP-Phone-Eval* dataset

feature sets	R
PART (Maximum R (LDA-Classifier (Maximum R (SVM (*MFCCs*), *PronFexP*)), *PronFexW*, *ProsFeat*))	0.61 (**)
PART (Maximum R (LDA-Classifier (Maximum R (SVM (*MFCCs*), *PronFexP*)), *PronFexW*, *ProsFeat*)), *RecAcc*	0.73 (**)
PART (Maximum R (LDA-Classifier (Maximum R (SVM (*MFCCs*), *PronFexP*)), *PronFexW*, *ProsFeat*)), *RecAcc*, *2-D Sammon Coordinates*	0.74 (**)
PART (Maximum R (LDA-Classifier (Maximum R (SVM (*MFCCs*), *PronFexP*)), *PronFexW*, *ProsFeat*)), *RecAcc*, *3-D Sammon Coordinates*	0.81 (**)
DecisionStump (Maximum R (LDA-Classifier (Maximum R (SVM (*MFCCs*), *PronFexP*)), *PronFexW*, *ProsFeat*))	0.57 (*)
SVM (SVM (SVM (*MFCCs*)))	0.41 (*)
SVM (SVM (SVM (*MFCCs*))), PART (Maximum R (LDA-Classifier (Maximum R (SVM (*MFCCs*), *PronFexP*)), *PronFexW*, *ProsFeat*)), *RecAcc*, *3-D Sammon Coordinates*	**0.82** (**)

Table 6.34: Correlations on speaker level between the percentage of detected words and the annotated percentage for weakened plosives on the *CLP-Phone-Eval* dataset. (*) marks significant correlations with $p < 0.05$ and (**) correlations with $p < 0.01$.

Experimental Setup

As mentioned in Chapter 4.1.3, the preprocessing is only relevant for phoneme and word level features as they have to be segmented properly. After segmentation, a decision has to be made whether the spoken word is children's or adults' speech.

Since the word chain to be uttered by the child is not known a priori, segmentation is much more difficult than in read speech where the reference is known. In order to improve the segmentation, a speech recognizer with a trigram language model is used. The language model is trained using the transcription of the speech tests of 262 children (namely the transcripts of the *Michael-Poeschke-06*, *CLP-02-06*, and *Preschool* databases). The categories of the language model were the 97 distinct words of the speech test used plus an additional category for words which appear in the "carrier sentences". In order to enable the recognition of unknown words, an out-of-vocabulary (OOV) word was added to each category. During this procedure several points have to be considered:

- **Correspondence of the spoken words to the test words:** Since the speech data were transliterated according to the acoustic realization of the child, the correspondence between the spoken words and the test words is not always clear. This is caused by the use of synonyms and pronunciation errors. In order to solve this problem, an alignment was performed between the transliteration of each test and the correct sequence of words with dynamic time warping (DTW) [Sanko 83]. In order to improve the alignment of pronunciation variants, the distance of substitutions was calculated according to the Levenshtein distance of the two words divided by the number of letters in the longer word. The procedure still has the problem that it is not capable to model variations in the sequence of the words which happen when a child names the words from right to left instead from left to right. Therefore, all found correspondences were checked manually according to their plausibility. Unplausible correspondences were removed. So about 20 different realizations of each word of the test were found. The most common correspondences can be reviewed in Appendix A.1.2.

- **Frequency of carrier words:** In the transliteration of the 262 speech test sessions two tendencies could be seen: Some children use many "carrier words" while others use none at all. Therefore, the segmentation is performed using two language models for each turn. A "big" one trained on sentences with two or more carrier words per slide, and a "small" one with two or less carrier words. Furthermore, two turn-dependent trigram language models were created. Again, one with two or more "carrier words" and one with two or less "carrier words" per target word. In preliminary experiments, trigram language models proved to yield the best recognition rates in all four cases compared to language models with larger or smaller context [Bockl 07a].

- **Recognition of unknown words:** To estimate the probability of the OOV words, each word which was observed less than three times was used to train the OOV language model probabilities. The probabilities of the OOV words in the language model were estimated using the VOCSIM algorithm [Gallw 02]. The acoustic realization of the OOV words is flat, e.g. it is assumed to be any sequence of the phonemes of the speech recognizer. The threshold for the VOCSIM algorithms for the creation of OOVs was chosen to be three as described in [Bockl 07a].

- **Compensation of age effects:** Several recognizers were trained for certain age groups. As previously evaluated in [Bockl 07a], the best groups for the creation of age-dependent recognizers were found as:

  - $<$ 7 years
  - 7 years
  - 8 years
  - 9+10 years
  - $>$ 10 years

  The adaptation procedure was performed on the acoustic models as well as on the HMM output probabilities as described in [Maier 05a, Maier 06d, Maier 08b].

- **System combination:** The recognition was performed for each turn using four different recognizers with the different language models as described above. In order to obtain a single word chain, the four best word chains plus the reference chain, i.e., the object names as shown on the slides, were merged using the Recognizer Output Voting Error Reduction (ROVER) [Fiscu 97, Maier 05b, Maier 05c, Maier 05d, Maier 08a].

In this manner, an improved recognized word chain is obtained. In [Bockl 07a], preliminary experiments were performed using the same data as in this work. An increase of the recognition rate of normal children speech from 64.7 % to 74.5 % WA was found. In the CLP speech data, this improvement was even better. The WA of -11.0 % of the baseline system without any adaptation was pushed to 42.6 %.

Another crucial point in the automatic processing is the identification of the speech data uttered by the speech therapist who recorded the speech data. After identification and segmentation into PLAKSS words, the data is analyzed with speaker identification techniques. Surprisingly, standard speaker identification methods were outperformed by a simple energy thresholding algorithm because the speech of the child is always louder than the speech of the therapist due to the head-mounted microphone. So, a class-wise averaged recognition rate of 96.5 % could be obtained. The recall for the class “children” was 98 % [Bockl 07a].

With the segmentation and the identification solved automatically, the experimental setup as displayed in Figure 6.5 is applied.

Results

The automatic identification and segmentation procedure could extract 2793 of the 2981 marked words and assign them successfully to a PLAKSS word. Of the 127 words which were marked as nasal by both raters, 113 were correctly segmented. In the opinion of the author, this segmentation is reliable enough for the further processing and classification.

	union		intersect	
classifier	CL	RR	CL	RR
OneR	50.0 %	94.0 %	50.2 %	**98.8 %**
DecisionStump	46.2 %	46.6 %	51.0 %	80.9 %
NaiveBayes	50.1 %	70.2 %	51.9 %	91.4 %
J48	47.0 %	64.4 %	51.6 %	84.7 %
PART	48.0 %	64.6 %	**52.6 %**	83.4 %
SVM	47.9 %	66.2 %	51.0 %	86.4 %

Table 6.35: Overview on the CL and RR of nasalization obtained by different classifiers on the *CLP-Phone-Eval2* dataset on frame level: The columns "union" and "intersect" denote the decision rule which was chosen in order to assign the labels.

In order to obtain a nasality label per phoneme, one can either decide for the union or the intersect of the perceptive ratings. In case of the union, the label "nasal" is assigned if one of the two raters found nasality. The intersect label is assigned if both raters agreed on their decision on the label "nasality". Both assignment rules were investigated. In case of the "union" rule, 449 nasal and 2344 non-nasal words were found. The "intersect" rule produced 113 nasal words and 2680 non-nasal words.

First evaluations on frame level showed that the label assignment rule "intersect" yielded more consistent labels than the "union" rule. Therefore, we chose for the "intersect" rule in the following experiments to obtain the labels.

From frame to phoneme level, most CL rates increase while the RR stays in the same range. Addition of the *TEP* to the *MFCCs* also improves the CL in most cases. *PronFexP* also succeed in improving the recognition. The best CL on phoneme level is obtained by a combination of all three feature types with 64.8 % CL.

As observed with the semi-automatically evaluated data, the recognition rates drop slightly when moving from phoneme to word level. Table 6.37 list the results obtained by the different classifiers with different feature sets. The results of unsuccessful combinations are not reported in the table. The best CL on word level is found with the NaiveBayes classifier with 62.1 % CL.

A receiver operated characteristics (ROC) evaluation [Fawce 06] of the NaiveBayes (NaiveBayes (NaiveBayes (*MFCCs*))) classifier shows a detailed report on the classification performance (cf. Figure 6.6). The axes denote the true and false positive rate, i.e., the trade-off between the percentage of correctly as "nasal" classified instances p(hit) and the percentage of false alarms p(false alarm). At a true positive rate of about 40 %, the number of false alarms is about 10 %. However, with increasing true positive rate, the number of false alarms also grows. At more than 65 % true positive rate, the number of false positives grows at the same rate: The classifier converges to the random classifier. Hence, the classifier can detect only about 40 % of the positives reliably. Note that the true positive rate of human rater 1 was 45.5 % at a false positive rate of 7.5 %, and rater 2 had a true positive rate of 61.5 % with a false negative rate of 5.7 % (cf. Table 6.6). The classifier is already close to rater 1, but rater 2 is still quite a lot better.

Although the classification on word level seemed rather weak, there are significant correlations between the classification results and the human evaluation. A combina-

classifier	feature sets	CL	RR
OneR	OneR (*MFCCs*)	49.9 %	**98.7 %**
DecisionStump	DecisionStump (*MFCCs*)	**62.4 %**	71.2 %
LDA-Classifier	LDA-Classifier (*MFCCs*)	57.5 %	84.5 %
NaiveBayes	NaiveBayes (*MFCCs*)	56.8 %	59.6 %
J48	J48 (*MFCCs*)	57.2 %	86.7 %
PART	PART (*MFCCs*)	**61.2 %**	80.3 %
SVM	SVM (*MFCCs*)	54.4 %	84.8 %
OneR	OneR (*MFCCs*), *TEP*	50.0 %	**98.8 %**
DecisionStump	DecisionStump (*MFCCs*), *TEP*	**63.9 %**	71.7 %
NaiveBayes	NaiveBayes (*MFCCs*), *TEP*	57.4 %	60.0 %
J48	J48 (*MFCCs*), *TEP*	56.5 %	87.0 %
PART	PART (*MFCCs*), *TEP*	**62.0 %**	80.0 %
SVM	SVM (*MFCCs*), *TEP*	55.1 %	85.4 %
OneR	PART (*MFCCs*), *TEP*, *PronFexP*	49.9 %	**98.7 %**
DecisionStump	PART (*MFCCs*), *TEP*, *PronFexP*	55.0 %	73.6 %
NaiveBayes	PART (*MFCCs*), *TEP*, *PronFexP*	54.3 %	74.4 %
J48	PART (*MFCCs*), *TEP*, *PronFexP*	**63.8 %**	82.7 %
PART	PART (*MFCCs*), *TEP*, *PronFexP*	60.8 %	82.4 %
SVM	PART (*MFCCs*), *TEP*, *PronFexP*	56.2 %	75.9 %
J48	PART (*MFCCs*), *TEP*, *PronFexP*	**64.8 %**	83.0 %

Table 6.36: Recognition rates on phoneme level for nasalization on the *CLP-Phone-Eval2* dataset

classifier	feature sets	CL	RR
OneR	OneR (OneR (*MFCCs*))	49.7 %	**94.0 %**
NaiveBayes	NaiveBayes (NaiveBayes (*MFCCs*))	**62.1 %**	75.8 %
PART	PART (PART (*MFCCs*))	54.7 %	65.6 %
SVM	SVM (SVM (*MFCCs*))	49.6 %	56.7 %
LDA-Classifier	LDA-Classifier (LDA-Classifier (*MFCCs*))	52.7 %	71.8 %
NaiveBayes	NaiveBayes (NaiveBayes (*MFCCs*), *TEP*)	**60.2 %**	81.8 %
SVM	SVM (PART (*MFCCs*), *TEP*, *PronFexP*)	57.7 %	67.2 %
SVM	SVM (PART (*MFCCs*), *TEP*, *PronFexP*), *PronFexW*	59.1 %	68.0 %
DecisionStump	Maximum R(AdaBoost (DecisionStump (Maximum R (PART (*MFCCs*), *TEP*))))	**58.4 %**	68.6 %
NaiveBayes	Maximum R(AdaBoost (NaiveBayes (Maximum R (PART (*MFCCs*), *TEP*))))	**61.4 %**	56.1 %
SVM	Maximum R(AdaBoost (SVM (Maximum R (PART (*MFCCs*), *TEP*, *PronFexP*))))	**59.7 %**	68.6 %

Table 6.37: Recognition rates on word level for nasalization on the *CLP-Phone-Eval2* dataset

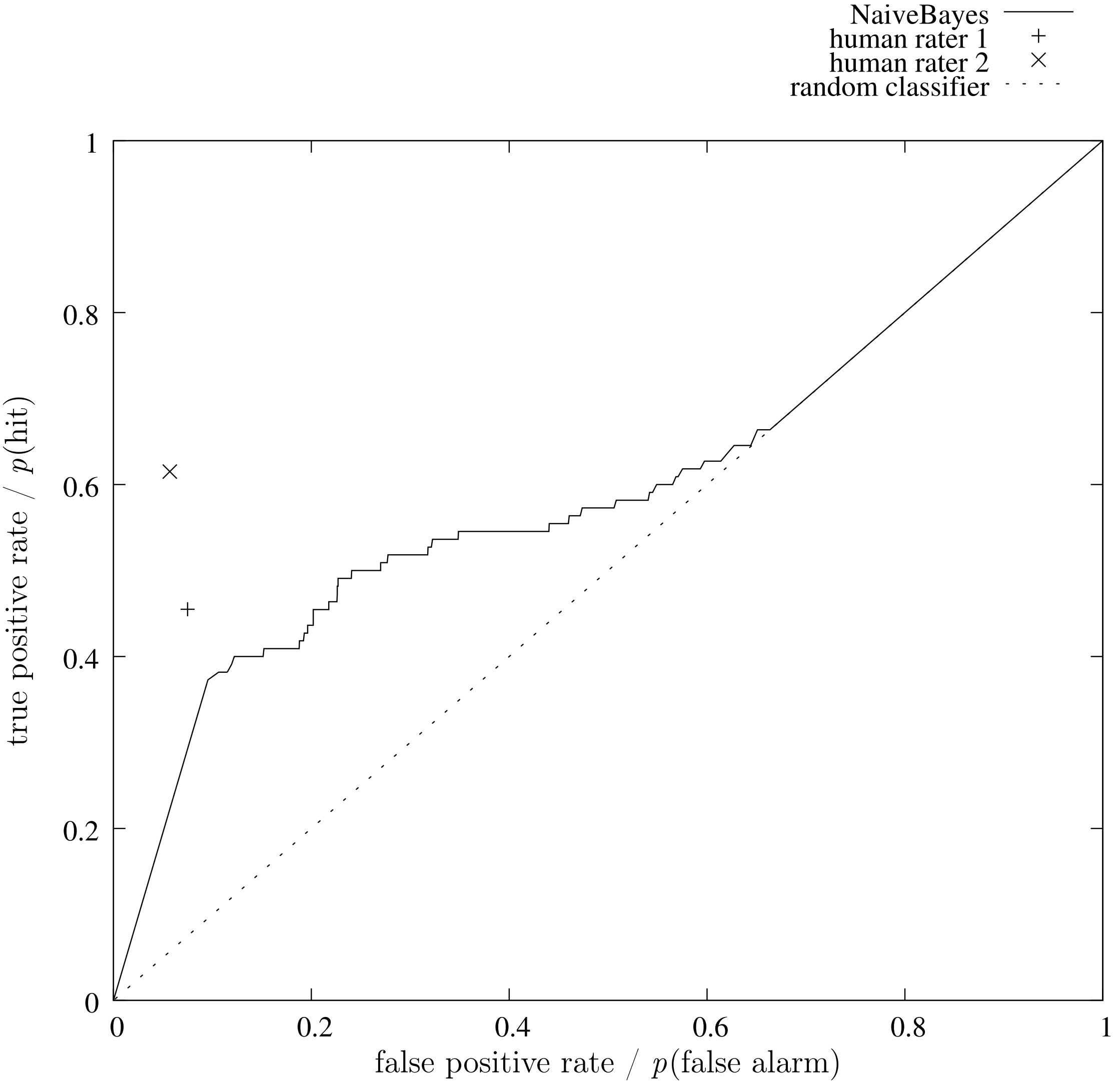


Figure 6.6: The ROC evaluation shows the strong and weak points of the NaiveBayes classifier: The CL of the classifier was 62.1 %. If the true positive rate and the false positive rate of the classifier are compared, it can be seen that the classifier works optimally at a true positive rate of about 30 %. Then, the false negative rate is less than 10 %. If more than 60 % true positive rate are required, the classifier converges to the random classifier, i.e., an increase in true positive rate brings the same increase of the false positive rate.

feature sets	R
SVM (SVM (PART (*MFCCs*), *TEP*, *PronFexP*), *PronFexW*)	0.35 (*)
NaiveBayes (Maximum R(AdaBoost (NaiveBayes (Maximum R (PART (*MFCCs*), *TEP*)))))	0.43 (**)
SVM (Maximum R(AdaBoost (SVM (Maximum R (PART (*MFCCs*), *TEP*, *PronFexP*)))))	0.45 (**)
SVM (SVM (PART (*MFCCs*), *TEP*, *PronFexP*), *PronFexW*)	0.49 (**)
SVM (SVM (PART (*MFCCs*), *TEP*, *PronFexP*), *PronFexW*), *2-D Sammon Coordinates*	0.61 (**)
SVM (SVM (PART (*MFCCs*), *TEP*, *PronFexP*), *PronFexW*), *3-D Sammon Coordinates*	0.63 (*)
SVM (SVM (PART (*MFCCs*), *TEP*, *PronFexP*), *PronFexW*), SVM (Maximum R(AdaBoost (SVM (Maximum R (PART (*MFCCs*), *TEP*, *PronFexP*))))), SVM (SVM (PART (*MFCCs*), *TEP*, *PronFexP*), *PronFexW*), *RecAcc*, *2-D Sammon Coordinates*, *3-D Sammon Coordinates*	0.81 (**)

Table 6.38: Correlations on speaker level between the percentage of detected words and the annotated percentage for nasality on the *CLP-Phone-Eval2* dataset. (*) marks significant correlations with $p < 0.05$ and (**) correlations with $p < 0.01$

tion of multiple information sources such as multiple classifiers on word level, features obtained from a Sammon map, and the *RecAcc* features yields a regression of up to 0.81. This result is significant at $p < 0.01$.

Evaluation on word and speaker level showed that the current system is able to judge the patients reliably on speaker level. The results on word level, however, still need a lot of improvement. Especially, the high false negative rates which are connected to high true positive rates are a problem. If the current classifier would be used to build an automatic therapy system, the acceptance would be low due to the high number of false alarms.

6.3.3 Discussion of the Results

In this section many results for the semi-automatic and automatic assessment of articulation disorders were presented. Some observations were consistent and some were not.

In general, RandomForests and OneR classifiers rather learn the original distribution of the data than the adjusted distributions. This results in high RRs because the majority class is recognized well. The recognition of the more rare class, however, is only weak.

The SVMs and tree-based classifiers yielded the best performance. Combination of classifiers with AdaBoost can improve these results in some but not in all cases.

Surprisingly, MFCCs alone yield high recognition rates. In most cases, MFCCs only already yield high recognition rates. We relate this to the fact that MFCCs model human perception of speech well in general. Hence, the effect of articulation disorders can also be seen in the MFCCs.

The use of functionals to raise the classification output from one level to the next higher level is very useful. From frame to phoneme level, the recognition virtually always increased. On word level, the phoneme level functionals also contributed to the recognition.

Combination of multiple features is beneficial on all evaluation levels. Especially, the pronunciation features in all articulation disorders and the *TEP* in the disorders concerning nasality. Hence, the pronunciation features can not only model the pronunciation errors by non-natives, but also articulation disorders in children. The *TEP* which was previously only used in vowels and consonant-vowel combinations showed to be applicable to connected speech as well. On speaker level, *RecAcc* and Sammon coordinates increased the correlation to the perceptive evaluation. This is in agreement with the factor analysis of the perceptual evaluation which showed that all articulation disorders are related to the speech intelligibility and hence also to *RecAcc* features.

Prosodic features performed weak in general. In most cases they did not contribute to any improvement. Reasons for this were already presented in Chapter 6.2.2.

Feature selection is beneficial in some cases, especially if many features are involved. It cannot be guaranteed that feature selection will improve the recognition.

The performance on frame, phoneme, and word level was rather weak in all experiments. Comparison of the automatic classification on word level with the human raters for the criterion “nasality” using a ROC evaluation showed that the automatic evaluation system is already close to one of the two raters. The other rater, however, performs much better than the automatic evaluation system. Hence, evaluation on word level is not as reliable as human raters yet.

Since only one rater was available for the semi-automatic assessment, no inter-rater correlations can be provided for this evaluation. For the fully automatic evaluation, however, an inter-rater correlation of 0.80 could be observed. The correlation between both raters and the automatic system was 0.81. Hence, the inter-rater correlation and the evaluation of the automatic system lie in the same range, i.e., both are equally reliable.

6.4 Visualization of the Data

Visualization of speakers and speaker dependencies can provide a better understanding of the speech disorders. However, the visualization or map of the speakers has to be meaningful, i.e., the quality has to be measured. In our case we decided to use three measurements for the evaluation:

- Sammon Error ϵ_{S}: The remaining error computed by the Sammon error function according to Eq. 4.73. This error is used to describe the loss of the mapping from the high-dimensional space to the low-dimensional space. In the literature, this term was shown to be a crucial factor to describe the quality of a representation [Shoza 04, Nagin 05, Hader 06b]. Since the scaling of the maps influences the Sammon error a lot, all maps were scaled in order to match their average Euclidean distances with the average distances of the high-dimensional data.

- Grouping Error e_{Grp}: The average distance between stars belonging to the same group (on a map with normalized coordinates in an interval between 0 and 1), i.e., the average distance between a speaker and his representation recorded with a different microphone.

  $$\epsilon_{\mathrm{Grp}} = \frac{1}{N} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \theta_{ij} g_{ij} \tag{6.1}$$

  Note that the normalization is just performed with $\frac{1}{N}$ due to the sparsity of $\boldsymbol{G}$. A grouping error of 0.25 corresponds to an average distance of 25 % of the maximum distance in the map between the representations of the same speaker.

- Regression: The regression between the coordinates of a map and a given criterion also provides information on the quality of the map. The regression is computed as described in Chapter 4.2.1.

In the following only maps computed with the Sammon mapping or the extended Sammon mapping are presented since both methods are superior to other dimensionality reduction methods like PCA or LDA [Exner 07].

As already shown in the previous section, the information of a visualization is also beneficial for the prediction of certain speaker properties (cf. Chapter 6.3).

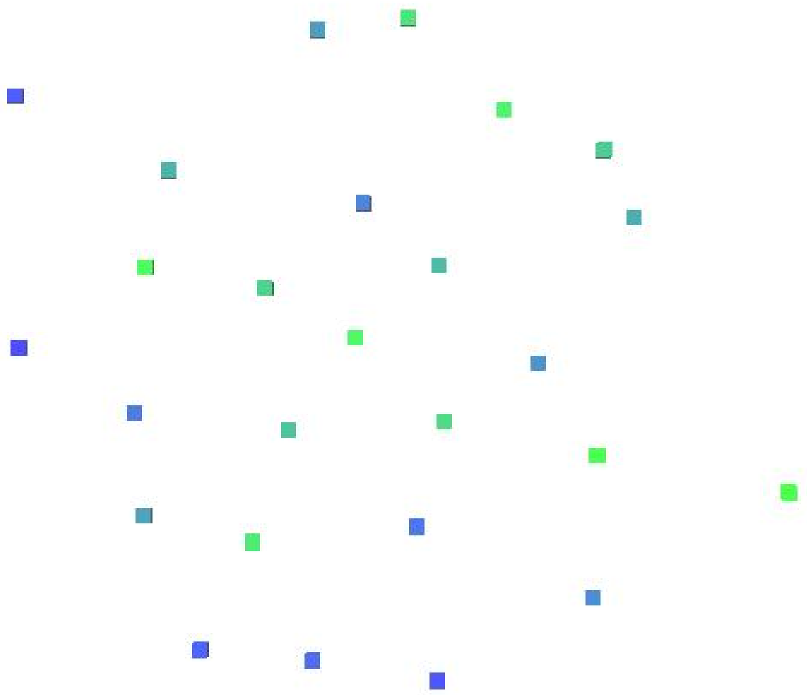


Figure 6.7: Visualization of the *CLP-Phone-Eval* data: The Intelligibility is coded in color; the brighter the color the more intelligible is the speaker. Regression between the coordinates and the different perceptively rated criteria are listed in Table 6.39 as “single”.

Figure 6.7 shows the speakers of the *CLP-Phone-Eval* database computed with the Sammon mapping as described in Chapter 4.2.3. The intelligibility (human ratings) is color-coded in the figure. The brighter the color the more intelligible is the speaker. The regression of the coordinates of the map to the intelligibility is 0.52 which is significant with $p < 0.05$. The coordinates also represent other criteria well (cf. Table 6.39, column “single”).

A visualization makes only sense if it is also robust to different recording conditions, especially in a client-server environment. As already described in Chapter 5, much of the data of this work was collected at different locations using different microphones. Most of the control data was recorded with a Plantronics Audio USB 510 headset while virtually all of the patient data was collected with a dnt Call 4U Comfort microphone.

In order to test whether our visualization method is independent of the microphone, we played the data of the *CLP-Intel* database back and recorded them with the Plantronics Audio USB 510 headset a second time. Furthermore, the data of the *Michael-Poeschke-07* database was also re-recorded with the dnt Call 4U Comfort microphone, i.e., both databases were re-recorded with the respective other microphone. For the playback the same Quadral SAM38A loud speaker was used as for the experiments in [Riedh 06]. In a robust visualization, the speakers should be projected onto the same or at least a close position in the map although two different microphones were used. Figure 6.8 (a) shows that this is not true for the usual Sammon mapping. While the patient group at the bottom of the map is projected to the same area of the map, the control group is split into two clusters which represent the two microphones. The acoustical mismatch between both versions of the control group is greater than the speaker characteristics, so the microphones dominate the distances and therefore also the clusters. In the patient group, the speaker characteristics are stronger than the acoustical mismatch caused by the microphone. Hence, both versions of the patient group from only one cluster.

In order to force the same speakers to be projected to the same location, the Sammon mapping is extended as described in Chapter 4.2.3. With the weight w_S, a trade-off between grouping and normal Sammon mapping is created. As Figure 6.8 shows, the points representing the same speaker move together with growing w_S.

Figure 6.9 shows the development of the grouping and the Sammon error in dependency of w_S. The higher w_S, the lower is the group error. The Sammon error increases with growing w_S. At $w_\mathrm{S} = 0.9$ a configuration is found where the sum of Sammon and grouping error is minimal as displayed in Figure 6.8 (c). $w_\mathrm{S} = 0.9$ seems to put most of the weight on the grouping error. However, if we recall the definition of ϵ_SE from Eq. 4.79 and the definition of $\mathcal{G}$ from Eq. 4.78, one can easily see that most of the error sum is caused by the Sammon error and not by the grouping error since $\mathcal{G}$ contains only N times an entry with $g_{i,j} = 1$ and $(N^2 - N)$ times $g_{i,j} = 0$. So if the average error would be equal, i.e.,

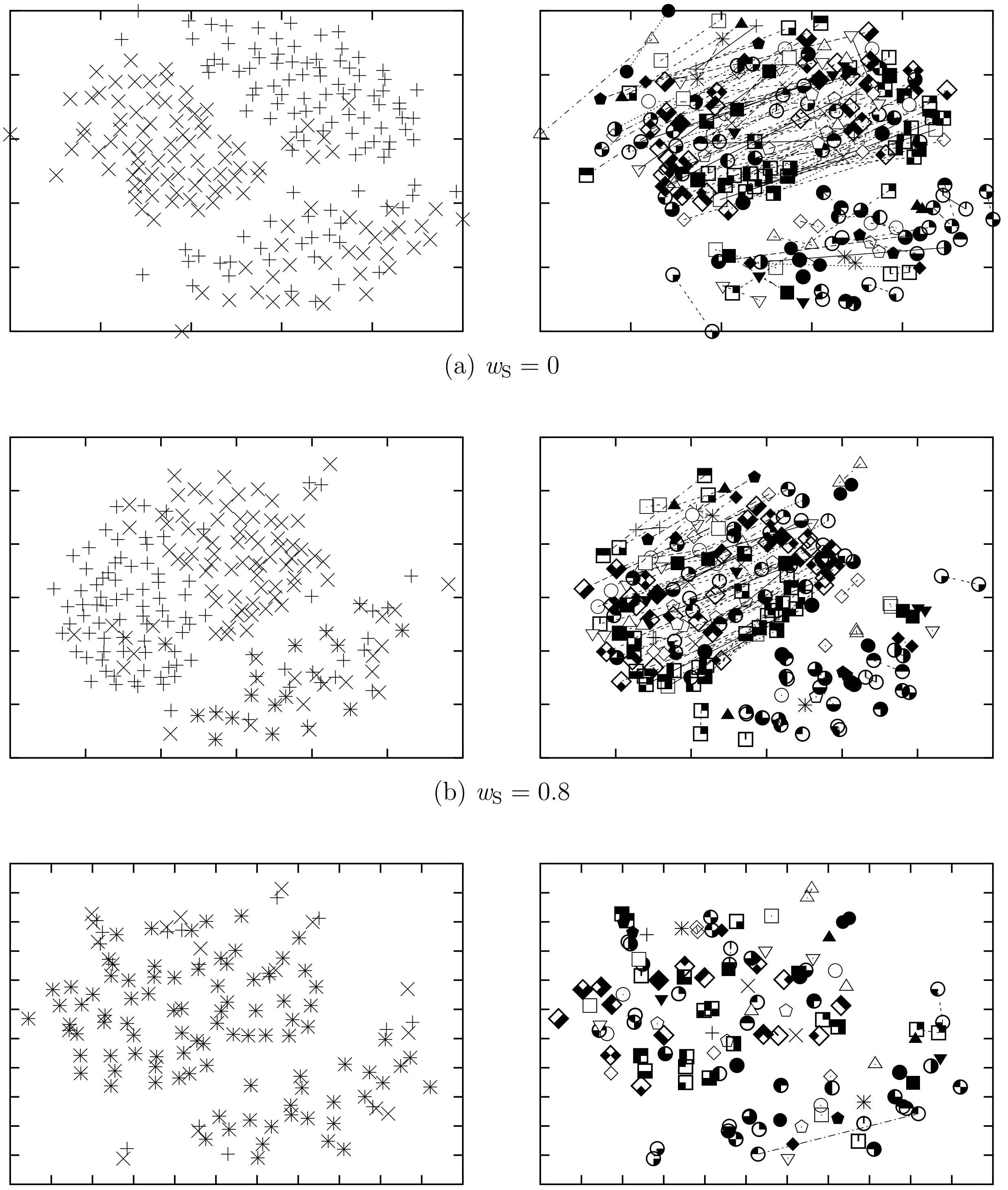
$$\sum_{p=1}^{N-1} \sum_{q=p+1}^{N} \frac{(\delta_{pq} - \theta_{pq})^2}{\delta_{pq}} \approx \sum_{p=1}^{N-1} \sum_{q=p+1}^{N} \theta_{pq} \tag{6.2}$$

the break-even point between both errors with $N \approx 200$ would be at about $w_\mathrm{S} = 0.99$. Hence, with $w_\mathrm{S} = 0.9$, the influence of the Sammon information is still very high.

In Table 6.39 correlations obtained with just a single microphone ("single") at $w_\mathrm{S} = 0$ and $w_\mathrm{S} = 0.9$ for the *CLP-Phone-Eval* subset of the visualization are reported. On the map which was created using just a single microphone, most correlations are significant or highly significant. However, if data using a different microphone is added, the correlations decrease a lot. The non-rigid registration helps to increase correlations again. Most of the correlations are in the same range as on the map created using only a single microphone. Hence, one can conclude that the registration of the different recording conditions yields robust and meaningful maps.

6.5 Summary

In this chapter the experiments and the results of this work were described. First the results of the perceptive evaluation of the speech intelligibility were presented. On the manually transliterated data (*CLP-Intel*), the inter-rater correlations were very consistent. The correlations were in the range from 0.92 to 0.96 (cf. Table 6.1). The raters of the automatically segmented data (*CLP-Intel2*) were also in very good agreement (cf. Table 6.3). The minimal inter-rater correlation was 0.87 while the



(a) $w_{\mathrm{S}} = 0$

(b) $w_{\mathrm{S}} = 0.8$

(c) $w_{\mathrm{S}} = 0.9$

Figure 6.8: Extended Sammon mapping on data played back with two different microphones: On the left side, the same microphone is marked with the same symbol ("X" for the Plantronics Audio USB 510 and "+" for the dnt Call 4U Comfort microphone). Each microphone forms a cluster although exactly the same speech data is represented. On the right side, each speaker is denoted with a unique symbol. The points which represent the same speaker are connected with a line: The shorter the lines, the fewer the grouping error. With growing w_{S} the grouping error is reduced. With $w_{\mathrm{S}} = 0.9$ almost no lines are visible, i.e., the grouping error is close to zero.

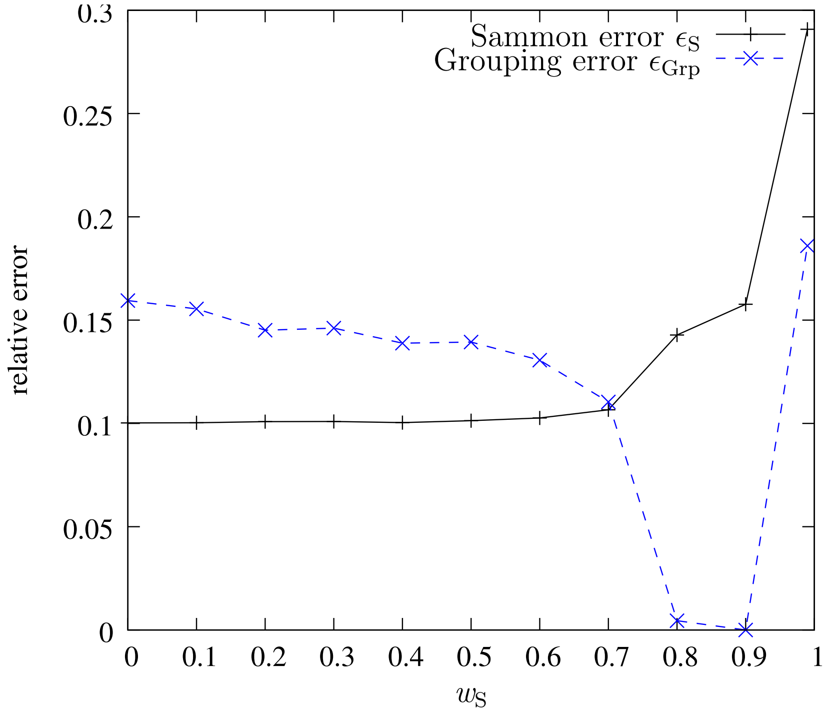


Figure 6.9: Development of the Sammon and the grouping error in dependency of w_S: While the Sammon error increases steadily with growing w_S, the grouping error decreases. With a too high weight of the grouping error, the coordinates become mere random numbers due to the random initialization.

criterion	single	$w_S = 0$	$w_S = 0.9$
nasalized consonants (NC)	0.31	0.10	0.42 (*)
laryngeal replacement (LR)	0.54 (*)	0.41 (*)	0.43 (*)
pharyngeal backing (PB)	0.61 (**)	0.39 (*)	0.43 (*)
weakened plosives (WP)	0.21	0.39 (*)	0.44 (*)
intelligibility	0.52 (*)	0.12	0.55 (**)
marked words	0.28	0.25	0.60 (**)
age	0.32	0.44 (*)	0.39 (*)

Table 6.39: Correlations on a Sammon map with only the patient group (“single”), in non-registered ($w_S = 0$), and registered ($w_S = 0.9$) conditions with multiple microphones. All measures are computed on the *CLP-Phone-Eval* subset of the visualization. (*) marks significant correlations at $p < 0.05$ while (**) marks highly significant correlations at $p < 0.01$.

semi-automatic evaluation							
	frame		phoneme		word		speaker
criterion	CL	RR	CL	RR	CL	RR	r
HN	56.8 %	99.0 %	62.9 %	99.0 %	60.6 %	96.9 %	0.89
NC	62.0 %	94.2 %	68.5 %	95.6 %	63.6 %	82.5 %	0.85
LR	59.8 %	99.6 %	69.5 %	99.6 %	63.8 %	98.2 %	0.81
PB	66.0 %	99.1 %	76.9 %	99.6 %	67.9 %	98.2 %	0.70
WP	71.1 %	97.8 %	71.1 %	97.8 %	75.8 %	97.8 %	0.82

fully automatic evaluation							
	frame		phoneme		word		speaker
criterion	CL	RR	CL	RR	CL	RR	r
nasalization	52.6 %	98.8 %	64.8 %	98.8 %	62.1 %	94.0 %	0.81

Table 6.40: Overview on the results of the pronunciation assessment

maximal was 0.95. On a subset of the *CLP-Intel* data (*CLP-Phone-Eval*), the pronunciation was assessed by an experienced speech therapist. With factor analysis, three components — a “nasalization”, a “backing”, and a “lisp” component — were identified. Furthermore, only the aspects with correlated highly with the “nasalization” and the “backing” components were significantly correlated with the intelligibility (HN, NC, LR, PB, and WP). The *CLP-Phone-Eval2* data were perceptively scored by two speech therapists according to the criterion “nasalization”. The inter-rater correlation was 0.80.

With the semi-automatic evaluation procedure, a correlation of up to -0.89 between the recognition result and the perceptive evaluation was obtained. With the fully automatic evaluation, a correlation between the experts and the automatic system of -0.93 was found. Further addition of prosody could improve only the semi-automatic system. Furthermore, the system showed significant differences between normal children and children with CLP. Normative values for control groups which were gathered in five cities in Germany were also provided. Application of speech quality measures, like 3SQM, were not able to distinguish normal from pathologic children.

The pronunciation assessment was performed with the *CLP-Phone-Eval* data for the criteria which were in significant correlation to the intelligibility. Moreover, the *CLP-Phone-Eval2* data were fully automatically evaluated for the criterion “nasalization”. Table 6.40 lists the best results which were found on the respective level.

The visualizations created with the Sammon mapping showed significant correlations to most of the perceptively rated criteria on the *CLP-Phone-Eval* data. Addition of further patient data recorded with a different microphone reduced these correlations a lot. An extension of the Sammon mapping, however, was able to restore the good correlations again.

Chapter 7

Outlook

Soon after the first version of PEAKS was finished, it became clear how powerful and handy the recording of speech data over the Internet is. While in the first few weeks only data of children with CLP were collected, the medical staff more and more realized how easy the use of the software is. More studies collecting data in different parts of the University Hospital of Erlangen were planned and carried out. The software ran on most PCs without any modifications. Sometimes, Java had to be installed or updated. The feedback of the medical personal is very positive. In the opinion of the doctors, PEAKS and other speech recognition techniques will simplify speech evaluation for clinical and scientific purposes.

By now, PEAKS is being used by four clinics in Erlangen, one in Herzogenaurach, and three clinics in Japan. The number of medical studies investigating the intelligibility of patients pro- and retrospectively is growing every month. At present, the complete database of the PEAKS server contains already approximately 5000 patient and control persons. Future extensions will concern refinement of the analyses, development of therapy tools, screening of speech, the investigation of more disorders, the integration of additional modalities and the internationalization of PEAKS.

Since PEAKS allows for an automatic evaluation of speech, it has a great advantage compared to the usual perceptive evaluations since they are not as reliable. Especially in outcome studies — in relation to evidence-based medicine — PEAKS shows great advantages because it can evaluate the global speech outcome parameter for all language, voice, and articulation disorders — the speech intelligibility. Currently, only the intelligibility of patients is assessed, using PEAKS in a large scale. In the future, the pronunciation evaluation as presented in this work will be used to investigate patients' speech in more detail. At speaker level, the evaluation is already so reliable that it can be applied in medical studies although the recognition on lower levels still contained errors. The errors, however, were consistent. Hence, they could be evened out on speaker level which is another advantage compared to perceptive evaluations. For the reliable evaluation on word and phoneme level, however, more research is required. In this work a lot of features which were already applied to second language learning were used. The results on word and phoneme level were in the same range as the results in the literature concerning pronunciation training. If further improvement is achieved, the software can also be applied for speech tutor systems.

Therapy sessions are expensive due to the fact that speech experts are rare and expensive. Hence, therapy could be enhanced by automatic sessions led by an automatic dialogue system. The dialogue can present explanations and exercises via speech synthesis and a graphical front-end. All input which is relevant to control the dialogue, however, must be performed using interactive input methods such as buttons and input dialogues since speech recognition accuracy is reduced in the patients. Such a kind of system could be administered by the speech therapist to a patient in order to ensure regular exercises. The exercises can then be performed independently of the therapist. After one or two weeks, the therapist can check whether the patient did the exercises regularly, and automatic summarization will give him an idea of the improvement the patient made within the training sessions.

Since PEAKS is also available as an offline version, it is ideal for the screening of speech. PEAKSlocal allows for the screening in kindergartens, schools, and at compulsory examinations of children. In this work, mean and standard deviations for children in school age all over Germany are given. However, for preschool screening more data of children in kindergarten age are required. Therefore, more data will be collected in the future. After recording of a whole kindergarten, parents and children with significantly low recognition rates could be invited to a speech therapist for further investigations.

Automatic evaluation of speech intelligibility is also interesting for all surgeons who perform changes to any part of the vocal tract. Currently, the German Cancer Aid is funding a project to investigate the speech of patient with partial removal of the larynx. PEAKS is also used in this project. Moreover, the Otolaryngology Department of the University Hospital of Erlangen researches the influence of structural changes of the nose and the nasal cavity on the intelligibility.

In the Pediatric Psychology Department of the University Hospital Erlangen, a group of scientists is also working on future applications of medical speech processing. In one project the reading ability of children is automatically determined. Again, the automatic speech processing should not replace any therapist but provide a reliable means of diagnosis. First results indicate that this is also possible. We investigated a group of children with reading disorders and one without reading disorders and annotated the number of reading errors manually. Significant correlations of up to 0.8 could be identified.

Investigation of other disorders, like language development disorders, is also interesting. While the feature sets and tests of the current version of PEAKS focus very much on the pronunciation evaluation, further modifications should allow the assessment of such language disorders. We suspect that the use of text mining techniques are promising for the analysis of language disorders. Bag-of-words features and language modeling techniques could be useful in such an automatic analysis.

Children with cochlea implants are able to gain a normal speech development today. The children understand as much as normal children. However, the implant reduces the number of tones which can be perceived. Hence, the children often have problems with the correct use of prosody. Even more difficult is the language acquisition in tonal languages, e.g. Chinese. A mispronunciation of the right tone might change the meaning of a word or sentence. Application of automatic evaluation techniques will provide feedback to the patients in order to learn the correct intonation.

Dementia is a disease which is related to a decline in neural processing ability. If the brain is not trained regularly, the risk of dementia is even higher. Therefore, automatic speech processing techniques will be useful to create a training system for the elderly. Such a system can be integrated into games or puzzles in order to increase the brain activity and to decrease the risk of dementia. However, the design must be matched to the user group since too complex systems might reduce the user comfort. Again, the therapist is not replaced. Instead, such puzzles and games should be regarded as new tools which are offered to the therapist. These tools might be able to extend and improve therapy because they offer exercises which can be performed additionally independent of the therapist.

As additional modalities, several new acquisition devices will be integrated into PEAKS. Using eye tracking technologies, for example, the visual scope of a patient can be acquired in a non-invasive manner. Modern eye tracking systems are integrated into normal LCD displays. Hence, the patient doesn't even realize that his visual focus of attention is being recorded. Eye tracking information can support the decoding of a speech recognizer since the visual focus should preceed the words which are said in tests like the PLAKSS test. Furthermore, the visual processing is related to reading disorders. Hence, certain visual perception patterns can be identified in children whose reading ability is delayed.

Even more challenging will be the automatic evaluation in children with fear disorders. Presentation of objects the child is afraid of should result in an avoidance reflex in the visual perception of the child while the speech of the child should express disgust. Multi-modal analysis of these patterns will bring further insight on this kind of disorder.

In several projects the use of tele-medical methods for speech therapy is investigated. First results indicate that such tele-medical therapy is also successful to extend the normal therapy because the patients can perform additional exercises at home. In current state-of-the-art systems, the exercises of a patient are recorded and transmitted to a speech expert. Due to bandwidth reasons, just random segments of the exercises are transmitted and evaluated perceptively by a speech therapist. Additional modules of PEAKS, which are still to be implemented, will allow for such exercises. The therapist could then check whether all administered exercises were performed, and an automatic, statistical summary of the exercises could be presented to him.

Even the perceptively evaluated tele-medical systems still have a lot of disadvantages. Since the camera angle is fixed, the therapist may not move around the patient in order to investigate breathing and body pose of the patient. Problems in breathing and a slack body pose might indicate the causes of certain speech disorders. Modern 3-D camera systems will be applied to improve the therapy because they allow a real-time acquisition of 3-D surface data using time-of-flight technology. With a color image mapped onto the surface data, the speech therapist can move around the patient virtually and investigate body pose or breathing of the patient. Speech therapy can be even more refined if a 3-D display is connected to the system. Then the therapist has a 3-D impression of the patient which is even more realistic.

Using the 3-D information and the color image, the movements of the patient will be automatically analyzed to create a statistical summary of the therapy session. The geometry of the face, for example, will reveal and measure metrically the type and extent of a facial paresis which is also important information to the therapist.

All methods presented here are mostly independent of the language. Only a speech recognition engine of the respective language is required. Therefore, the translation of PEAKS into different languages is also possible. A Japanese version has already been created, and versions in English, Portuguese, Dutch, Italian, Czech, and Swedish are already work in progress. In the future this list will be expanded to more languages.

The application of automatic speech recognition techniques to medical problems is still a very young field of research. In fact, the first system which was able to assess the speech intelligibility was designed and developed at the University Erlangen-Nuremberg in 2006. The chances which lie in this field are overwhelming. And so is the number of applications and challenges that will be explored in the future.

Chapter 8

Summary

In this work the automatic evaluation of speech of children with cleft lip and palate was investigated. In the introduction a general definition of language, voice, and speech disorders was given. Language disorders have their origin right at the beginning of the speech production chain: the brain. Voice disorders appear in the excitation of the voice, i.e., at the vocal folds. Speech of children with CLP often contains articulation disorders. The disorders are caused by either anatomical changes in the structure of the vocal tract or — even after adequate treatment — misarticulations due to changed anatomy during the acquisition of speech.

Chapter 2 described cleft lip and palate in detail. First, the epidemiology, the embryology, and the functional consequences were reported. 80 % of all orofacial clefts include CLP. Its prevalence ranges from 1 in 400 to 500 newborns in Asians, 1 in 750 to 900 newborns in Caucasians, and 1 in 1500 to 2000 newborns in African Americans. The cleftings develop in the human embryo from week 7 to week 10. Functional consequences concern nutrition, swallowing, breathing, speech disorders, and hearing loss. The state-of-the-art treatment already begins in the first few hours after the birth of the child. A palatal obturator is placed in the mouth to close the clefting of the palate until the child is old enough for primary surgery. The first surgeries take place between the sixth and 15th month. Speech therapy begins according to the individual needs of the child. If required, further surgeries are performed between the 12th and 18th year of the child.

The state-of-the art evaluation of disordered speech was discussed in Chapter 3. Most evaluation methods rely on the perceptive evaluation of speech. However, all perceptive evaluation methods lack objectivity. A commonly used method to reduce subjectivity is the use of a panel of experts. This allows for a *inter-rater-confirmed-subjective-mean-score* which is often referred to as “objective”. Next, the state-of-the-art automatic evaluation methods were described. In the scope of this work, these methods were referred to as objective. Methods for the objective evaluation of nasality exist, but they are either invasive, like the Nasometer, or analyze only sustained vowels or consonant-vowel combinations but not connected speech. The only non-invasive method to assess speech intelligibility in connected speech was developed in Erlangen and is based on the evaluation of the speech data using a speech recognition system.

In Chapter 4 the system which was developed during this work was described. The overview on PEAKS stated the fundamental requirements which have to be fulfilled in order to be applicable in a medical environment. The two use cases for the system portrayed the tasks which are to be performed by the user and the administrator of the system. Functional requirements of the system were the evaluation of the speech intelligibility of patients, the automatic assessment of certain aspects of speech, and the visualization of speakers in order to allow their comparison. Non-functional requirements were multi-user support, platform-independency, security concepts, collaboration concepts, and user comfort.

PEAKS is built as a pattern recognition system. However, it is divided into a client and a server part. Only recording and presentation of the results is done at the client. All analyses are performed on the server side. These methods were described in Chapter 4.2. In order to evaluate the performance of the system and the human raters, measurements of agreement had to be discussed. The most common measures for the agreement of two raters are correlation coefficients. Therefore, Pearson's correlation, Spearman's rank correlation, and the Multiple Regression / Correlation Analysis were described. For correlation coefficients, significance tests are available. The measurement of agreement for more than two raters is commonly performed with Kappa and Alpha. Both, however, proved to be inappropriate if the evaluations result on different scales, e.g. comparisons between human evaluators and a speech recognition engine.

A major part of this work relied on regression and classification. However, not all used classifiers could be described in detail. Therefore, Support Vector Machines and Regression were chosen to be analyzed in detail. Both methods are based on the use of Support Vectors to model either the hyperplanes which separate the cases optimally or the regression function, i.e., the model is not determined explicitly. The model consists of a subset of the training data which is best suited to represent the model.

Another important point of this work is the reduction of dimensionality. It is on the one hand important for the reduction of feature spaces before classification and on the other hand necessary for the visualization in a low-dimensional space. First, linear methods for the dimensionality reduction were introduced. All linear methods are based on a matrix multiplication to achieve the reduced dimensionality. The Principal Components Analysis finds the transformation matrix which projects the data onto the dimension with the highest variance. The Linear Discriminant Analysis emphasizes the components which separate the data best. Feature selection can also be interpreted as a linear dimensionality reduction. Here, the dimensions are kept unchanged. However, one has to choose a selection criterion. A new method for the efficient calculation of the Multiple Regression / Correlation was presented. As a popular method for the non-linear dimension reduction, the Sammon mapping was explained. Furthermore, an extension of the Sammon mapping to create microphone-independent maps was introduced.

Speech processing and speech recognition are very important parts in order to understand PEAKS. Speech recognition was explained from the basics starting from the common Mel Frequency Cepstrum Coefficients which were used as features in the PEAKS speech recognition system. Another important method for speaker identification and speech recognition are Gaussian Mixture Models (GMMs). Classification and training were described in detail. GMMs are also used in the acoustic modeling of the speech recognizer. The acoustic models consist of Hidden Markov Models (HMMs) which are able to model the dynamic properties of speech. In our case the HMM states are tied with individual weights to a single GMM with 500 output densities. The acoustic model is combined with a statistical language model. During the recognition procedure, both are used to decode the best, i.e., the most likely, word chain.

Based on this word chain, more features can be computed from the speech data. Prosodic features extract pitch and speaking manner from the signal using fundamental frequency, energy, and duration features. Pronunciation features are computed from the phone confusion probabilities between the recognized word chain and the reference word chain. Furthermore, output probability features, as the "Goodness Of Pronunciation", and recognition accuracy features are part of the pronunciation feature set. Moreover, a special feature to detect hypernasality in vowels — the Teager Energy feature — was described.

In order to compensate the effects of age on speech, adaptation techniques as Maximum Likelihood Linear Regression, and feature normalization techniques, as Vocal Tract Length Normalization, were explained.

The Architecture of PEAKS is divided into three blocks. The classes of the client are used to form the graphical user interface, the perceptive evaluation forms, and the recording environment. The transport classes handle the secure transmission of the data and state the transfer objects which hold all relevant information in the system. The server block offers a gateway to the SQL database to store transfer objects. Furthermore, the server classes wrap the speech recognition engine and provide wrapper functions for the decoding of the speech data and feature extraction.

Chapter 5 describes the speech data which were collected during this work. All children were recorded with the PLAKSS test. The test consists of 99 pictograms which are shown on 33 slides. Each of the slides tests a specific consonant at the beginning, in the center, and at the end. The test contains all German phonemes. A total of 1088 children were recorded during this work. 857 children form a control group which was recorded in five major cities in Germany.

Already starting from 2002, children with CLP were recorded in the Oral and Maxillofacial Clinic of the University Hospital Erlangen. Until 2006, 123 children were recorded. These data had to be transliterated in order to be able to perform automatic processing. Speech intelligibility was annotated by five speech experts for 31 children of these recordings. A full speech examination by an experienced speech therapist was done for 26 of these children. Furthermore, 50 data sets of children from a preschool study were transliterated. Using PEAKS, 189 CLP children were recorded between early 2006 and 2008. A subset of 35 children was scored with respect to the intelligibility. Moreover, the nasality of 32 children was evaluated by two speech therapists.

In Chapter 6 the results obtained with PEAKS were reported. The perceptive evaluation of the transliterated and non-transliterated data showed good consistency with correlations between 0.87 and 0.96. Factor analysis of the detailed speech examination showed three main factors. Two components — the "nasality" and the "backing" component — were in high correlation with the perceptual speech intelligibility while the third component — a "lisp" component — was uncorrelated to the intelligibility. Hence, the automatic evaluation was only performed for the criteria which were related to the speech intelligibility.

Automatic evaluation of speech intelligibility was first performed with the speech recognizer only. A correlation between the perceptive evaluation and the recognition accuracy of up to -0.89 was found on the transliterated data while a correlation of up to -0.93 was obtained with the fully automatic system. In case of the transliterated data, further addition of prosodic features could refine the prediction of the experts' scores. In both cases the evaluations of the automatic systems were in the same range as the human experts.

The automatic evaluation of the criteria concerning speech intelligibility on frame, phone, word, and speaker level was also investigated. In order to obtain an optimal automatic processing, the results of each level were combined using functionals to form features which are then added to the next level of the assessment procedure. Due to the sparsity of the misarticulated events, a high recognition rate was found for all criteria on all levels. The class-wise averaged recognition rate (CL), however, showed, that the recognition of most criteria was only moderate on frame, phone, and word level. But these moderate results were combined with more features on speaker level to yield a robust prediction at an average correlation of 0.82 ± 0.06.

The fully automatic evaluation of nasality was also performed on frame, phone, word, and speaker level. Again, the results on frame, phone, and word level were moderate. The combination on speaker level, however, proved to be as reliable as in the semi-automatic case. The quality of the automatic system on speaker level was in the same range as the speech experts with a correlation of 0.81 of the automatic system and an inter-rater correlation of 0.80 of the human raters.

Visualization of the speech data was also successfully performed. The coordinates of the projected speakers showed moderate to high correlations with the different perceptively evaluated criteria. However, the use of more than one microphone for the projection reduced this dependency. The meaning of the coordinates was restored mostly by the application of the extended Sammon mapping for multiple microphones.

PEAKS features the first fully automatic system in the world to assess speech intelligibility and articulation disorders in connected children's speech as reliable as human experts. In the future further extensions of PEAKS are planned. These will concern the refinement of the analyses, the investigation of additional disorders, the integration of further modalities, and the internationalization of PEAKS.

Appendix A

Details of the PLAKSS Test

On the following pages, further details on the PLAKSS test [Fox 02] are presented. The first two sections list the original test design and the target words (cf. Appendix A.1.1). Since some of the words are ambiguous, the vocabulary was extended by frequent word alternatives in order to enable their recognition. These are presented in Appendix A.1.2. Harcourt Test Services approved the use of the PLAKSS test for the scientific experiments in this work.

Subsequently the forms are presented which were used for the subjective evaluation (cf. Appendix A.2). The forms list all target words of the PLAKSS test. The target phones are underlined. For each word, scores can be given according to different criteria, like hypernasality, nasal air emission, and laryngealizations. Moreover, the forms for the parental approval of the recording are presented. Details on the pronunciation scores given by the speech expert on the databases are listed in Appendix A.4

In the last section of the chapter, some documentary images of the different recording setups are shown.

A.1 Vocabulary of the PLAKSS Test

A.1.1 Original Vocabulary

Slide 1:	Mond	Eimer	Baum
Slide 2:	Ball	Gabel	Blume
Slide 3:	Brief	Brille	Zebra
Slide 4:	Pilz	Wippe	Korb
Slide 5:	Pferd	Apfel	Topf
Slide 6:	Vogel	Marienkäfer	Schiff
Slide 7:	Pflaster	Flasche	Frosch, Quak
Slide 8:	Wurst	Löwe	Lampe
Slide 9:	Teller	Ball	Nuss
Slide 10:	Kanne	Telefon	Dusche
Slide 11:	Feder	Rad	Drachen
Slide 12:	Tasse	Auto	Bett
Slide 13:	Trecker	Zitrone	Jäger
Slide 14:	Milch	Eichhörnchen	Taucher
Slide 15:	Buch	Roller	Schere
Slide 16:	Gießkanne	Nagel	Berg
Slide 17:	Glas	Gras, grün	Schlange
Slide 18:	Kuh	Jacke	Sack
Slide 19:	Kleid	Krokodil	Knöpfe
Slide 20:	Sonne	Hase	Haus
Slide 21:	Zange	Katze	Pilz
Slide 22:	Zwerg	Hexe	
Slide 23:	Schuh	Tasche	Fisch
Slide 24:	Schlüssel	Schmetterling	Schnecke
Slide 25:	Spinne	Schrank	Schwein
Slide 26:	Stuhl	Kiste	Nest
Slide 27:	Spritze	Strumpf	Rutsche
Slide 28:	Anker	Bank	Punkt
Slide 29:	Arzt	Bild	Hund
Slide 30:	Fenster	Gespenst	Schornstein
Slide 31:	Erdbeere	Heizung	Elefant
Slide 32:	springt	kaputt	Unfall
Slide 33:	Tiger	Gitarre	

Table A.1: Original target words of the PLAKSS test [Fox 02, Bockl 07a]

A.1.2 Extended Vocabulary

Anker	Angel, Haken
Apfel	Äpfel
Arzt	Ärztin, Doktor, Mann
Auto	Auto
Ball	Ball, Baum, Rad, Wasserball, Zwiebel
Bank	Stuhl
Baum	Tannenbaum
Berg	Berg, Berge, Burg, Gebirge, Gipfel, Wiese, Zelt
Bett	Wiege
Bild	Berg, Bilderrahmen
Blume	Blumen, Sonnenblume
Brief	Bild, Brief, Briefmarke, Briefumschlag, Karte, Post, Postkarte, Schein
Brille	Sonnenbrille
Buch	Heft
Drachen	Drache, Drachensteigen
Dusche	Badewanne, Duschen
Eichhörnchen	Einhörnchen, Hörnchen
Erdbeere	Beere, Erdbeer, Erdbeeren
Feder	Vogel
Fenster	Gardinen, Vorhang
Flasche	Flaschen, Flaschenpost, Medizin, Saft, Wein
Gabel	Rechen
Gespenst	Geist, Geister, Gespenster, Mumie
Gießkanne	Gießer, Kanne
Glas	Glas, Tasse, Weinglas
Gras	Wies, Wiese, Rasen
grün	grünes
Hase	Hasen, Osterhase
Heizung	Heizer, Heizkörper
Hexe	Hexe, Zauberer
Hund	Wauwau
Jacke	Anorak, Anzug, Hemd, Mantel, Pulli, Pullover, Rock, T-Shirt
Jäger	Hund, Mann
Kanne	Flasche, Gießkanne, Kaffee, Kaffeekanne, Tasse, Teekanne, Vase, Wasser
kaputt	Auto, Autos, kaputte, kaputtes, Schrott, Tut
Katze	Kater, Katz, Mietzekatze, Tatze
Kiste	Karton, Kasten, Kissen, Koffer, Schachtel, Schubfach, Schublade
Kleid	Anziehen, Anzug, Hemd, Jacke, Kleidung, Mantel, Pullover, Rock, T-Shirt

Table A.2: Common word alternatives of the PLAKSS target words (A–Kl)

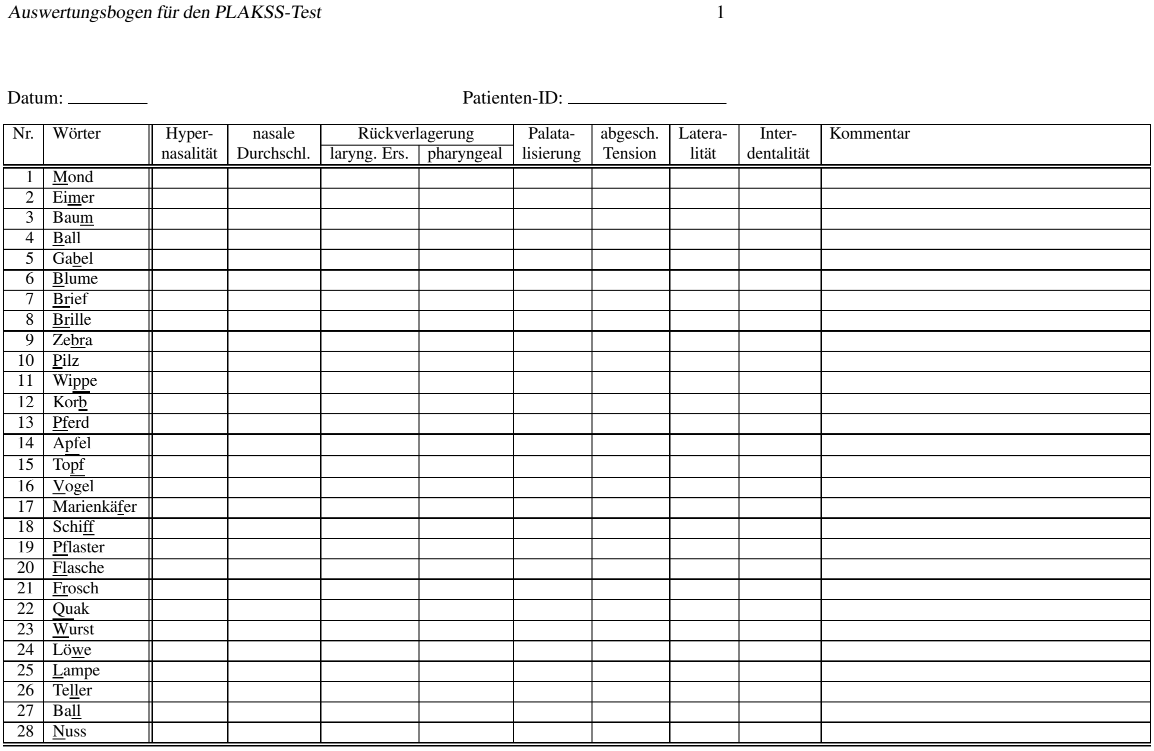
Knöpfe	Knopf, Knöpf, Knöpfe, Knöpfen
Korb	Eimer, Körbchen, Tor
Kuh	Ziege
Lampe	Glühbirne, Licht, Lichter
Löwe	Löwen, Tiger
Marienkäfer	Ameise, Junikäfer, Käfer, Maikäfer
Milch	Kuh, Melken, Milchkanne, Trinken
Mond	Halbmond
Nagel	Nadel, Nagen, Nager, Schraube
Nest	Eier, Netz, Taube, Vogel, Vogelnest
Nuss	Eichel, Haselnuss, Kastanie, Kokosnuss, Nüsse
Pferd	Pferdchen, Pony
Pflaster	Heftpflaster
Pilz	Fliegenpilz, Giftpilz, Pilze
Punkt	Ball, Fleck, Kugel, Murmel, Punkt, Stein
Quak	quaken
Rad	Fahrrad, Reifen
Roller	Fahrrad, Karren, Rad, Roller
Rutsche	Rumpf, Rutschbahn, Rutschen
Sack	Beutel, Jacke, Müll, Rucksack, Sand
Schiff	Boot, Pirat, Segelboot
Schlüssel	Schlüssel, Schüssel
Schmetterling	Marienkäfer, Schmetterling
Schnecke	Schneck, Schnelle, Versteck
Schornstein	Brief, Dach, Haus, Kamin, Ofen, Rauch, Schlot, Stein
Schrank	Kleiderschrank
Schuh	Schuhe, Stiefel
Schwein	Schein, Schweinchen, Wildschwein
Sonne	Sonnen
Spinne	Spinnen, Spinnennetz
springt	hüpfen, Hüpfseil, hüpft, Kind, Mensch, Seil, Seilhüpfen, Seilspringen, Seilspringer, seilspringt, spielt, Springe, springen, Springseil, Springseile, Sprung
Spritze	Spitze, Spitzer
Strumpf	Schuh, Socke, Socken, Stoff, Strümpfe, Strumpfhose

Table A.3: Common word alternatives of the PLAKSS target words (Kn–S)

Tasche	Geldbeutel, Handtasche
Tasse	Becher, Glas, Kaffeetasse, Kanne, Tee
Taucher	Mann, Mensch, Schwimmen, Schwimmer, tauchen, taucht, Wasser, Wassermann
Teller	Besteck, Gedeck, Messer
Tiger	Leopard, Löwe, Tier, Tiger
Topf	Kochtopf, Pfanne
Trecker	Bagger, Bulldock, Bulldog, Laster, Motor, Traktor
Unfall	Auto, Autos, Autounfall, Autozusammenstoss, Crash, Notfall, Strasse, überfahren, Verkehr, Zusammengestossen, Zusammenstoss
Vogel	Rabe, Vogel, Vögel
Wippe	Schaukel, Wippen
Wurst	Fleisch, Fleischwurst, Salami, Würstchen
Zange	Gabel, Kanne, Knacker, Nagel
Zebra	Esel, Zebras
Zitrone	Zitronen
Zwerg	Baby, Kasper, Kind, kleines Männchen, Männchen, Männlein, Sandmann, Wichtel, Zauberer

Table A.4: Common word alternatives of the PLAKSS target words (T-Z)

A.2 Forms for the Subjective Assessment of the PLAKSS Test

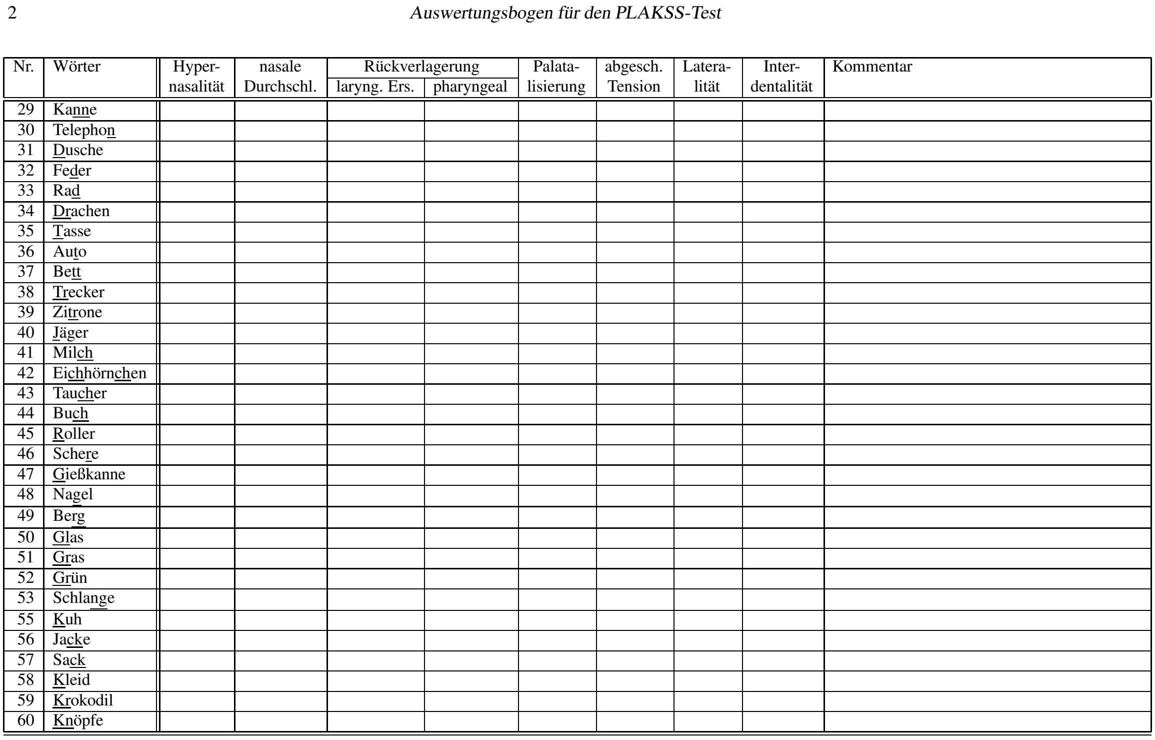


Auswertungsbogen für den PLAKSS-Test 1

Datum: ________ Patienten-ID: ________________

Nr.	Wörter	Hyper- nasalität	nasale Durchschl.	Rückverlagerung		Palata- lisierung	abgesch. Tension	Latera- lität	Inter- dentalität	Kommentar
				laryng. Ers.	pharyngeal					
1	Mond									
2	Eimer									
3	Baum									
4	Ball									
5	Gabel									
6	Blume									
7	Brief									
8	Brille									
9	Zebra									
10	Pilz									
11	Wippe									
12	Korb									
13	Pferd									
14	Apfel									
15	Topf									
16	Vogel									
17	Marienkäfer									
18	Schiff									
19	Pflaster									
20	Flasche									
21	Frosch									
22	Quak									
23	Wurst									
24	Löwe									
25	Lampe									
26	Teller									
27	Ball									
28	Nuss									

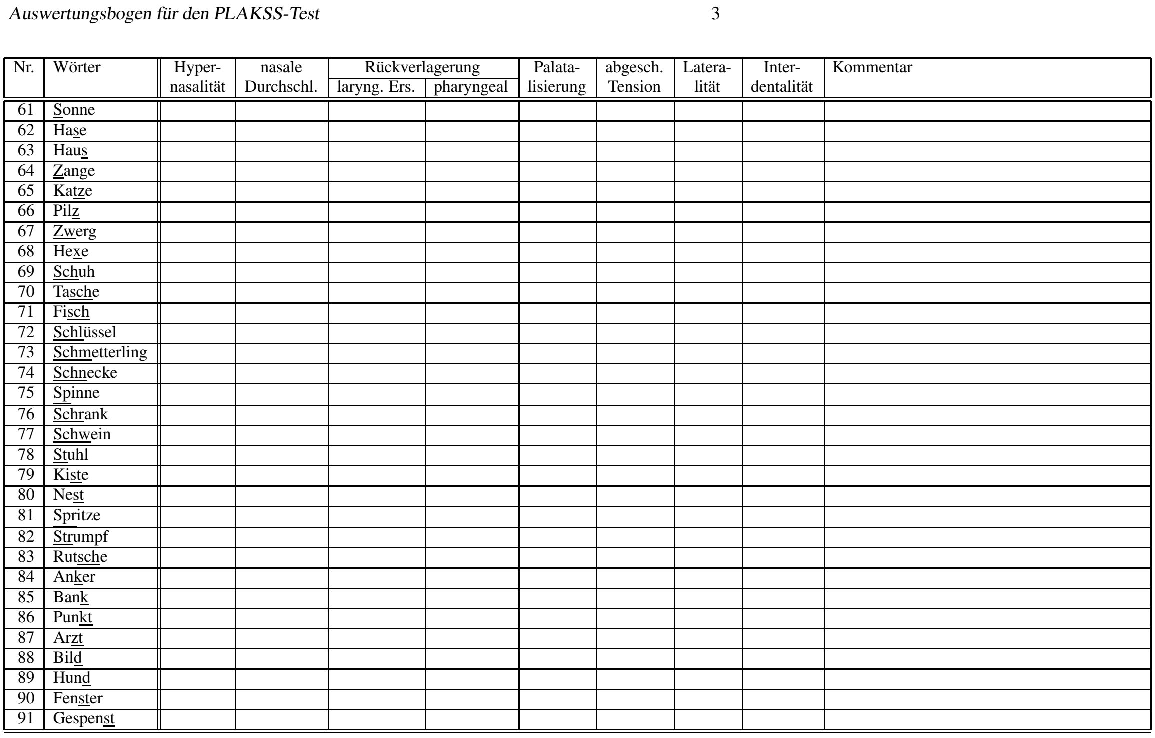
Figure A.1: Form used for the assessment of the PLAKSS test (page 1)



2 *Auswertungsbogen für den PLAKSS-Test*

Nr.	Wörter	Hypernasalität	nasale Durchschl.	Rückverlagerung		Palatalisierung	abgesch. Tension	Lateralität	Interdentalität	Kommentar
				laryng. Ers.	pharyngeal					
29	Kanne									
30	Telephon									
31	Dusche									
32	Feder									
33	Rad									
34	Drachen									
35	Tasse									
36	Auto									
37	Bett									
38	Trecker									
39	Zitrone									
40	Jäger									
41	Milch									
42	Eichhörnchen									
43	Taucher									
44	Buch									
45	Roller									
46	Schere									
47	Gießkanne									
48	Nagel									
49	Berg									
50	Glas									
51	Gras									
52	Grün									
53	Schlange									
55	Kuh									
56	Jacke									
57	Sack									
58	Kleid									
59	Krokodil									
60	Knöpfe									

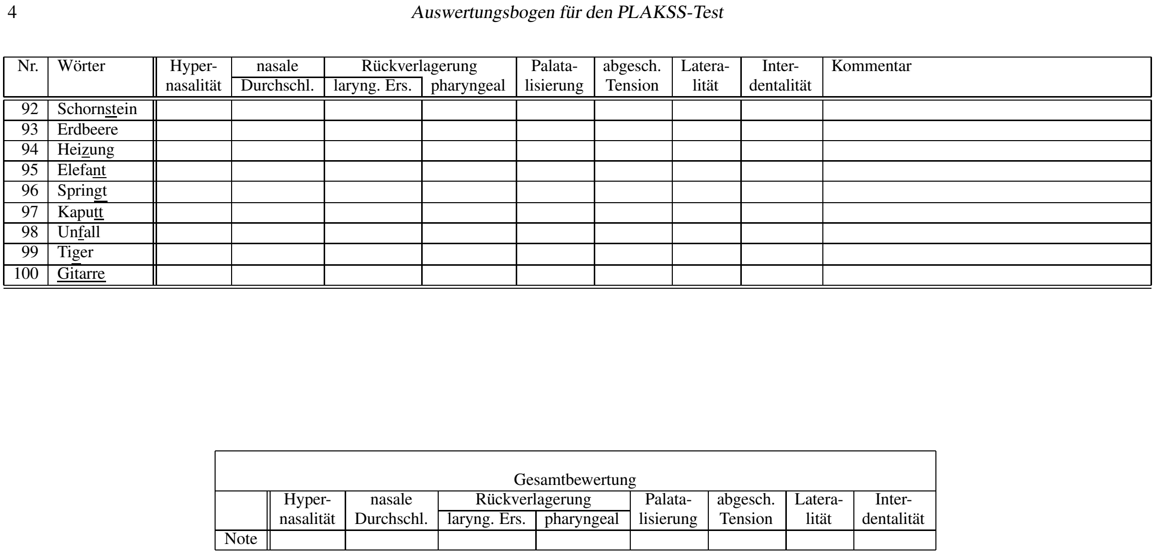
Figure A.2: Form used for the assessment of the PLAKSS test (page 2)



Auswertungsbogen für den PLAKSS-Test 3

Nr.	Wörter	Hypernasalität	nasale Durchschl.	Rückverlagerung		Palatalisierung	abgesch. Tension	Lateralität	Interdentalität	Kommentar
				laryng. Ers.	pharyngeal					
61	Sonne									
62	Hase									
63	Haus									
64	Zange									
65	Katze									
66	Pilz									
67	Zwerg									
68	Hexe									
69	Schuh									
70	Tasche									
71	Fisch									
72	Schlüssel									
73	Schmetterling									
74	Schnecke									
75	Spinne									
76	Schrank									
77	Schwein									
78	Stuhl									
79	Kiste									
80	Nest									
81	Spritze									
82	Strumpf									
83	Rutsche									
84	Anker									
85	Bank									
86	Punkt									
87	Arzt									
88	Bild									
89	Hund									
90	Fenster									
91	Gespenst									

Figure A.3: Form used for the assessment of the PLAKSS test (page 3)



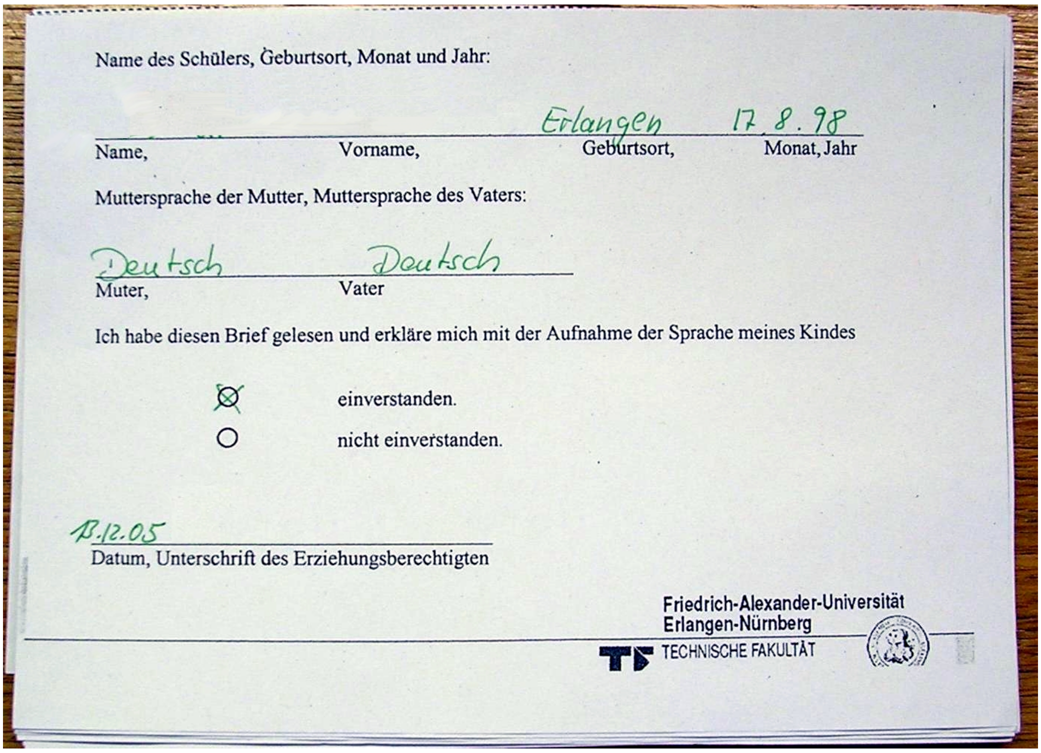
4 *Auswertungsbogen für den PLAKSS-Test*

Nr.	Wörter	Hypernasalität	nasale Durchschl.	Rückverlagerung		Palatalisierung	abgesch. Tension	Lateralität	Interdentalität	Kommentar
				laryng. Ers.	pharyngeal					
92	Schornstein									
93	Erdbeere									
94	Heizung									
95	Elefant									
96	Springt									
97	Kaputt									
98	Unfall									
99	Tiger									
100	Gitarre									

Gesamtbewertung

	Hypernasalität	nasale Durchschl.	Rückverlagerung		Palatalisierung	abgesch. Tension	Lateralität	Interdentalität
			laryng. Ers.	pharyngeal				
Note								

Figure A.4: Form used for the assessment of the PLAKSS test (page 4)



Name des Schülers, Geburtsort, Monat und Jahr:

Erlangen 17.8.98

Name, Vorname, Geburtsort, Monat, Jahr

Muttersprache der Mutter, Muttersprache des Vaters:

Deutsch Deutsch

Muter, Vater

Ich habe diesen Brief gelesen und erkläre mich mit der Aufnahme der Sprache meines Kindes

☒ einverstanden.

○ nicht einverstanden.

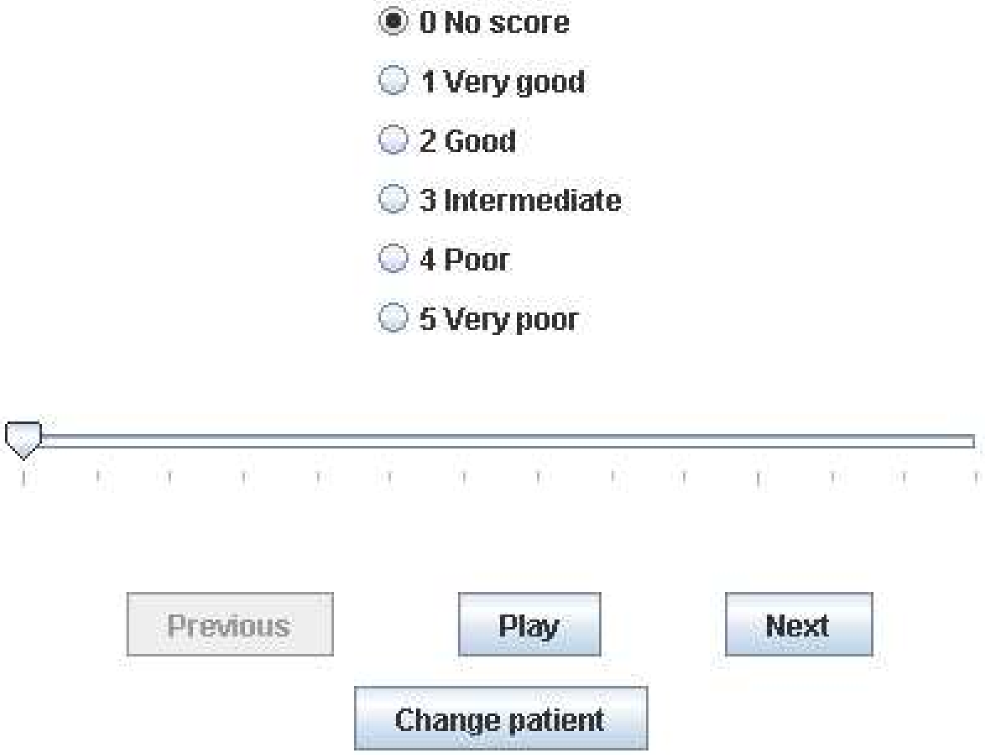
13.12.05

Datum, Unterschrift des Erziehungsberechtigten

Friedrich-Alexander-Universität
Erlangen-Nürnberg
TECHNISCHE FAKULTÄT

Figure A.5: Form for the parental approval to the recording (cf. Chapter 5.2)

A.3 Screenshots of the Subjective Evaluation Software



● 0 No score

○ 1 Very good

○ 2 Good

○ 3 Intermediate

○ 4 Poor

○ 5 Very poor

Figure A.6: Panel used for the subjective assessment of speech intelligibility (cf. Chapter 6.2)

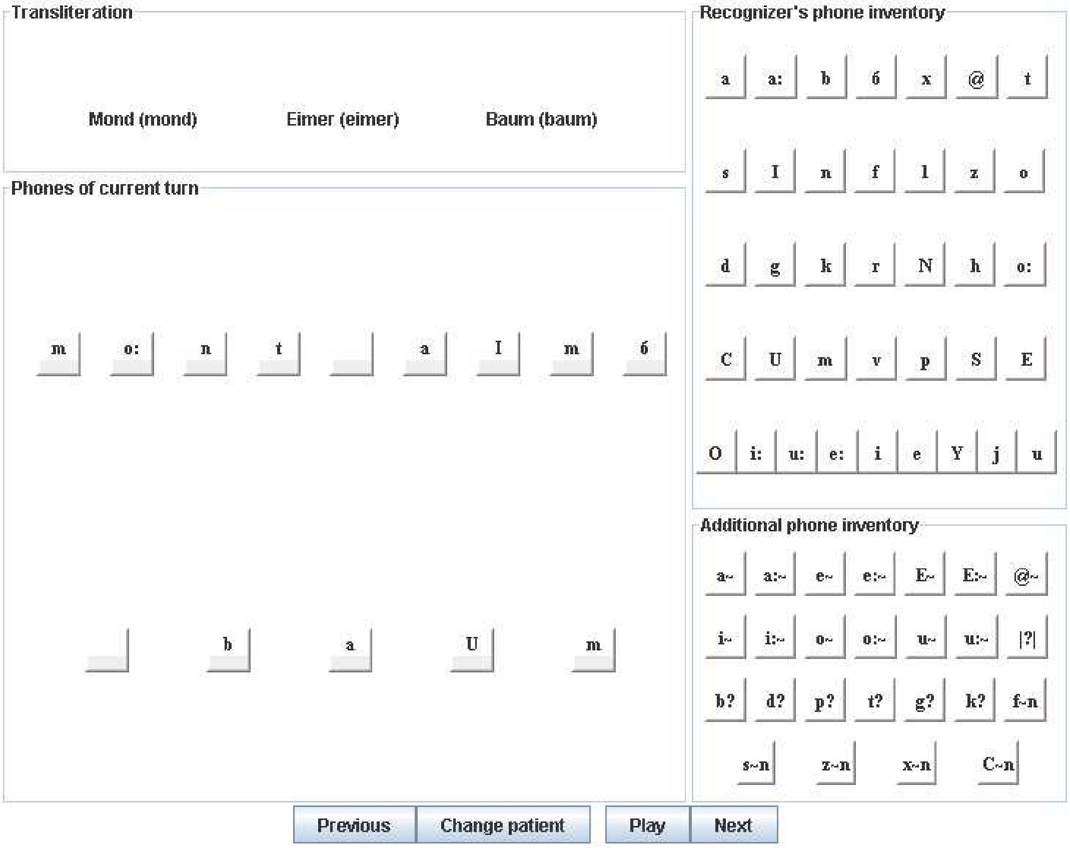


Figure A.7: Panel used for the subjective assessment of individual phones (cf. Chapter 6.2). Phones are denoted in SAMPA annotation while additional sounds which appear only in speech of children with CLP are denoted with diacritics. "~" denotes a realization with enhanced nasal air emission while "?" marks a glottal realization. "|?|" is a complete laryngeal realization similar to a glottal stop.

A.4 Subjective Annotation of the Data

A.4.1 Intelligibility Scores

patient ID	rater B	rater K	rater L	rater S	rater W
w010000s01	2.83	2.69	2.69	1.62	3.41
w010001f01	4.59	4.68	4.68	4.33	4.90
w010004f01	1.38	2.00	2.00	1.16	1.72
w010005f01	1.70	1.82	1.82	1.32	1.90
w010007f01	3.07	3.22	3.22	2.25	3.32
w010009f01	2.55	3.18	3.18	2.12	3.08
w010014f01	2.59	2.16	2.16	1.90	1.92
w010022f01	4.30	3.93	3.93	3.82	4.61
w010024f01	2.60	2.99	2.99	2.42	3.04
m010002f01	3.07	3.54	3.54	2.28	2.57
m010003f01	3.52	3.21	3.21	1.92	3.12
m010006f01	3.70	3.46	3.46	2.86	3.97
m010008f01	1.66	1.88	1.88	1.29	1.66
m010010f01	3.36	3.64	3.64	2.83	3.52
m010011f01	2.30	3.15	3.15	1.71	2.85
m010012f01	4.71	4.92	4.92	4.73	4.87
m010013f01	3.29	3.55	3.55	2.32	3.24
m010015f01	4.58	4.89	4.89	4.45	4.88
m010016f01	2.41	2.98	2.98	1.24	2.76
m010017f01	4.37	3.55	3.55	3.08	4.27
m010018f01	2.67	2.39	2.39	1.28	2.33
m010019f01	2.82	2.79	2.79	1.62	2.56
m010020f01	2.88	2.80	2.80	1.74	2.64
m010021f01	2.41	2.84	2.84	1.87	2.60
m010023f01	1.53	2.47	2.47	1.62	1.68
m010025f01	2.93	3.14	3.14	2.21	2.90
m010027f01	2.36	2.00	2.00	1.50	2.27
m010028f01	1.80	1.43	1.43	1.42	2.12
m010029f01	1.84	1.96	1.96	1.14	2.62
m010030f01	1.45	1.49	1.49	1.16	2.04
m010031f01	1.61	1.60	1.60	1.19	1.67

Table A.5: Experts' scores of the patients in the *CLP-Intel* database

patient ID	rater M	rater L	rater S	rater W
20000	2.23	2.23	2.00	2.87
20001	1.93	2.40	2.25	3.64
20002	1.00	1.33	1.55	2.30
20003	4.50	4.92	4.64	4.89
20004	2.30	2.80	2.30	2.61
20005	1.67	2.15	2.33	2.39
20006	4.86	4.67	4.56	4.95
20007	3.33	3.80	4.12	4.12
20008	3.00	3.46	3.17	2.94
20009	1.50	2.46	2.11	1.42
20010	1.00	1.31	1.83	1.00
20011	1.15	1.08	1.88	1.07
20012	2.00	2.42	2.32	1.83
20013	1.43	1.00	1.28	1.55
20014	2.09	2.42	2.55	2.76
20015	1.38	1.42	1.72	1.55
20016	1.44	1.33	1.45	1.61
20017	4.12	3.27	2.72	3.18
20018	1.54	1.50	2.06	1.88
20019	1.25	1.17	1.48	1.34
20020	1.47	1.08	1.54	1.19
20021	2.47	2.15	2.65	2.86
20022	1.07	1.00	1.64	1.12
20023	2.30	2.33	2.58	2.64
20025	3.26	2.85	3.52	3.94
20026	1.36	1.17	1.55	1.24
20027	3.38	3.36	3.22	4.06
20028	1.93	1.50	1.78	2.13
20029	3.06	1.33	2.37	2.36
20030	1.40	1.25	2.12	3.24
20031	4.56	4.91	4.84	5.00
20032	3.00	2.64	2.48	3.58
20033	1.39	1.68	1.83	2.30
20034	2.46	3.09	2.70	3.23
20035	2.39	1.19	2.14	2.55

Table A.6: Experts' scores of the patients in the *CLP-Intel2* database

A.4.2 Pronunciation Assessment

patient ID	HN	NC	LR	PB	WP	IN	LA	$\sum$
w010000s01	0	0	0	0	0	0	0	0
w010001f01	0	56	20	12	14	1	1	104
w010004f01	0	19	0	1	0	16	12	48
w010005f01	0	0	0	0	0	0	0	0
w010007f01	0	9	0	0	3	1	6	19
m010002f01	0	0	0	0	0	15	6	21
m010003f01	0	28	0	0	0	0	0	28
m010006f01	0	0	0	0	0	0	24	24
m010008f01	0	0	0	0	0	0	0	0
m010010f01	1	32	0	0	0	12	0	45
m010011f01	0	0	0	6	0	0	0	6
m010012f01	0	30	9	6	9	0	8	62
m010013f01	15	44	0	0	3	0	14	76
m010015f01	25	37	0	0	26	1	1	90
m010016f01	0	0	0	0	0	0	0	0
m010017f01	9	23	2	4	8	0	36	82
m010018f01	0	0	0	0	3	0	0	3
m010019f01	0	41	0	0	4	0	3	48
m010020f01	0	8	0	0	12	35	0	55
m010021f01	0	0	0	1	4	1	39	45
m010023f01	0	5	0	0	0	20	1	26
m010025f01	0	0	0	2	3	0	22	27
m010028f01	0	0	0	1	0	0	19	20
m010029f01	0	0	0	0	17	0	0	17
m010030f01	0	0	0	0	0	0	0	0
m010031f01	0	0	0	0	0	4	0	4
$\sum$	50	332	31	50	106	106	192	867

Table A.7: Pronunciation assessment on the *CLP-Intel* database

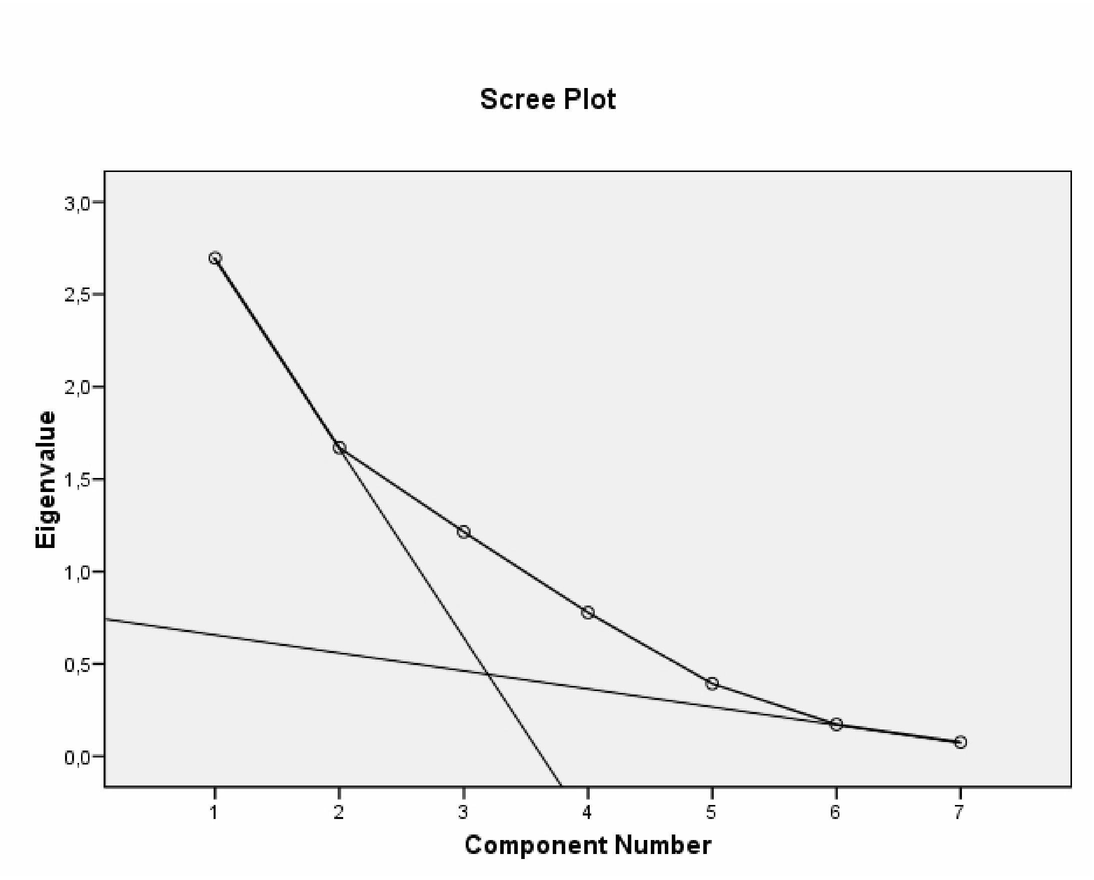


Figure A.8: Scree plot of the eigenvalues obtained by factor analysis on *CLP-Intel*

A.5 Documentary Images of the Recording Environment

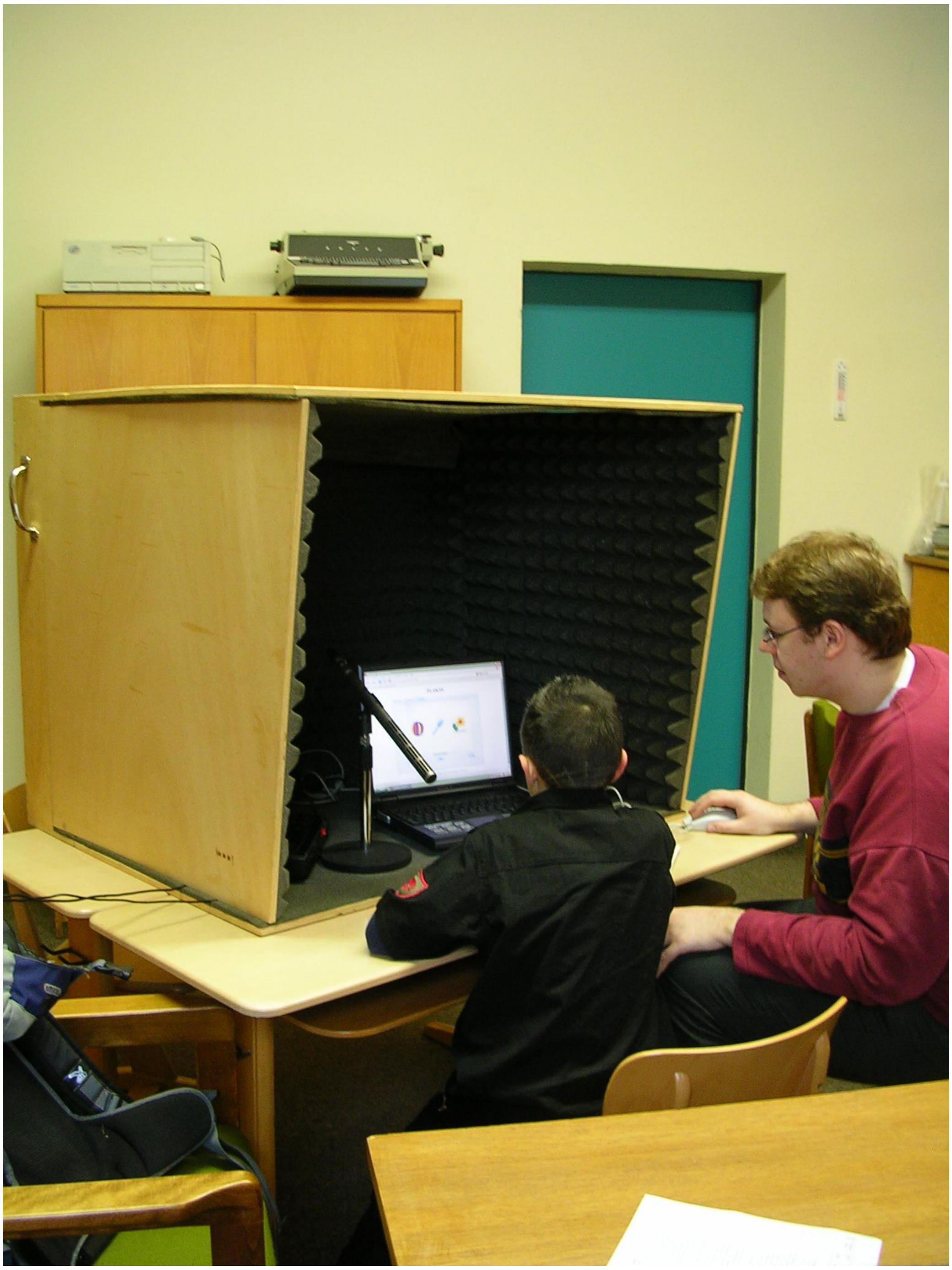


Figure A.9: Original recording environment used in 2006 for the recording of the first control group in a school in Erlangen

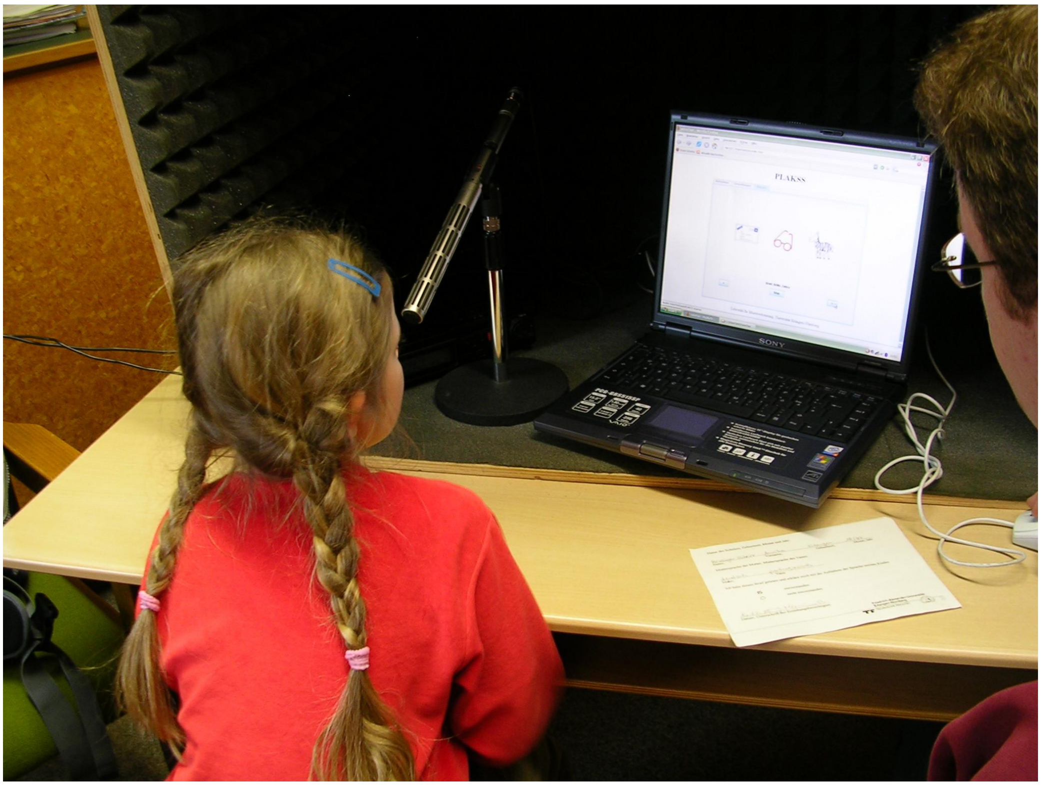


Figure A.10: Original recording environment used in 2006 (detail view)

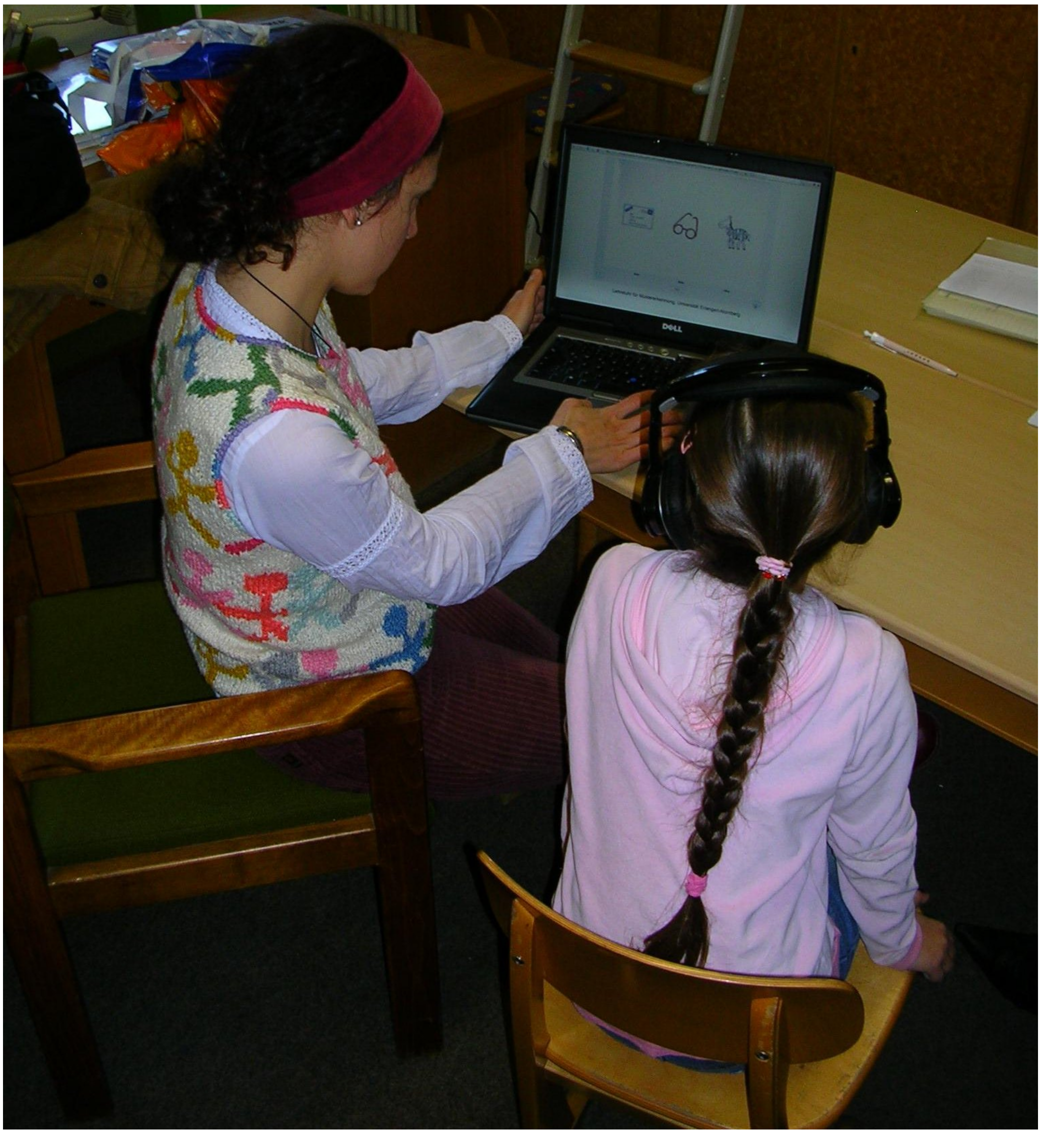


Figure A.11: Simplified recording environment used in 2007

List of Figures

List of Tables

Bibliography

[Adelh 03] J. Adelhardt, R. Shi, C. Frank, V. Zeißler, A. Batliner, E. Nöth, and H. Niemann. "Multimodal User State Recognition in a Modern Dialogue System". In: *the 26th German Conference on Artificial Intelligence*, pp. 591–605, Springer, 2003.

[Allis 06] B. Allison, D. Guthrie, and L. Guthrie. "Another Look at the Data Sparsity Problem". In: P. Sojka, I. Kopeček, and K. Pala, Eds., *9th International Conf. on Text, Speech and Dialogue (TSD)*, pp. 327–334, Springer, Berlin, Heidelberg, New York, 2006.

[Ander 73] M. Anderberg. *Cluster Analysis for Applications.* Academic Press, New York, USA, 1973.

[Atal 67] B. Atal and M. Schroeder. "Predictive Coding of Speech Signals". In: *Proc. Conf. Communication and Processing*, pp. 360–361, 1967.

[Atal 79] B. Atal and M. Schroeder. "Predictive coding of speech signals and subjective error criteria". *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 27, No. 3, pp. 247–254, 1979.

[Bagsh 93] P. Bagshaw, S. Hiller, and M. Jack. "Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching". In: *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pp. 1003–1006, ISCA, Berlin, Germany, 1993.

[Batli 00] A. Batliner, A. Buckow, H. Niemann, E. Nöth, and V. Warnke. "The Prosody Module". In: W. Wahlster, Ed., *Verbmobil: Foundations of Speech-to-Speech Translation*, pp. 106–121, Springer, New York, Berlin, 2000.

[Batli 01] A. Batliner, J. Buckow, R. Huber, V. Warnke, E. Nöth, and H. Niemann. "Boiling down Prosody for the Classification of Boundaries and Accents in German and English". In: *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pp. 2781–2784, ISCA, Aalborg, Denmark, 2001.

[Batli 03a] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. "How to Find Trouble in Communication". *Speech Communication*, Vol. 40, pp. 117–143, 2003.

[Batli 03b] A. Batliner, V. Zeissler, C. Frank, J. Adelhardt, R. P. Shi, and E. Nöth. "We are not amused - but how do you know? User states in a multimodal dialogue system". In: *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pp. 733–736, ISCA, Geneva, Switzerland, 2003.

[Batli 04] A. Batliner, C. Hacker, S. Steidl, E. Nöth, S. D'Arcy, M. Russell, and M. Wong. ""You stupid tin box" - children interacting with the AIBO robot: A cross-linguistic emotional speech corpus". In: *Proceedings of the 4th International Conference of Language Resources and Evaluation (LREC))*, pp. 171–174, 2004.

[Batli 95] A. Batliner, R. Kompe, A. Kießling, E. Nöth, H. Niemann, and U. Kilian. "The prosodic marking of phrase boundaries: Expectations and Results". In: A. Rubio, Ed., *New Advances and Trends in Speech Recognition and Coding*, pp. 325–328, Springer–Verlag, Berlin, 1995.

[Batli 99] A. Batliner, J. Buckow, R. Huber, V. Warnke, E. Nöth, and H. Niemann. "Prosodic Feature Evaluation: Brute Force or Well Designed?". In: *Proc. of the 14th Intl. Congress of Phonetic Sciences (ICPhS)*, pp. 2315–2318, San Francisco, USA, 1999.

[Bautz 08] W. Bautz. "Hompage of the Interdisciplinary Cleft Lip and Palate Center of the University Clinic Erlangen". 2008. http://www.lkg-zentrum.uk-erlangen.de, last visited 06/30/2008.

[Beere 02a] J. G. Beerendes, A. W. Rix, M. P. Hollier, and A. P. Hekstra. "Perceptual Evaluation of Speech Quality (PESQ) The New ITU Standard for End-to-End Speech Quality Assessment, Part I – Time-Delay Compensation". *J Audio Eng Soc*, Vol. 50, No. 10, 2002.

[Beere 02b] J. G. Beerendes, A. W. Rix, M. P. Hollier, and A. P. Hekstra. "Perceptual Evaluation of Speech Quality (PESQ) The New ITU Standard for End-to-End Speech Quality Assessment, Part II – Psychoacoustic Model". *J Audio Eng Soc*, Vol. 50, No. 10, 2002.

[Bockl 07a] T. Bocklet. "Optimization of a Speech Recognizer for Medical Studies on Children in Preschool and Primary School Age". Diplomarbeit, Chair of Pattern Recognition, University Erlangen–Nuremberg, Erlangen, Germany, 2007.

[Bockl 07b] T. Bocklet. "Speaker Recognition and Recognition of Speaker Groups Using Gaussian Mixture Models". Studienarbeit, Chair of Pattern Recognition, University Erlangen–Nuremberg, Erlangen, Germany, 2007.

[Bockl 07c] T. Bocklet, A. Maier, and E. Nöth. "Text-independent Speaker Identification using Temporal Patterns". In: V. Matoušek and P. Mautner, Eds., *10th International Conf. on Text, Speech and Dialogue (TSD)*, pp. 318–325, Springer, Berlin, Heidelberg, New York, 2007.

[Bockl 08a] T. Bocklet, A. Maier, J. Bauer, F. Burkhardt, and E. Nöth. "Age and Gender Recognition for Telephone Applications based on GMM Supervectors and Support Vector Machines". In: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1605–1608, IEEE Computer Society Press, Las Vegas, USA, 2008.

[Bockl 08b] T. Bocklet, A. Maier, and E. Nöth. "Age Determination of Children in Preschool and Primary School Age with GMM-based Supervectors and Support Vector Machines/Regression". In: P. Sojka, I. Kopeček, and K. Pala, Eds., *11th International Conf. on Text, Speech and Dialogue (TSD)*, pp. 253–260, Springer, Berlin, Heidelberg, New York, 2008.

[Bosel 04] M. Boseley and C. Hartnick. "Assessing the outcome of surgery to correct velopharyngeal insufficiency with the pediatric voice outcomes survey". *Int J Pediatr Otorhinolaryngol*, Vol. 68, pp. 1429–1433, 2004.

[Brand 05] H. Brandl. "Sprechermodellierung auf geringen Trainingsstichproben". Diplomarbeit, Ernst-Moritz-Arndt-University Greifswald, Germany, 2005.

[Breim 01] L. Breiman. "Random Forests". *Machine Learning*, Vol. 45, No. 1, pp. 5–32, 2001.

[Bress 02] T. Bressmann, R. Sader, P. Jürgens, H. F. Zeilhofer, and H. H. Horch. "Sprechsprachliche Ergebnisse nach einfachen und mehrfachen Gaumenverschlussoperationen". *Mund-, Kiefer-, Gesichtschirurgie*, Vol. 6, pp. 98–101, 2002.

[Bress 98] T. Bressmann, R. Sader, M. Merk, W. Ziegler, R. Busch, H. Zeilhofer, and H. Horch. "Perzeptive und apparative Untersuchung der Stimmqualität bei Patienten mit Lippen-Kiefer-Gaumenspalten". *Laryngorhinootologie*, Vol. 77, No. 12, pp. 700–708, 1998.

[Bress 99a] T. Bressmann, R. Sader, S. Awan, R. Busch, H.-F. Zeilhofer, and H.-H. Horch. "Quantitative assessment of hypernasality in patients with cleft lip and palate by computerised measurement of nasalance". *Mund-, Kiefer- und Gesichtschirurgie*, Vol. 3, No. 1, pp. 154–157, 1999.

[Bress 99b] T. Bressmann, R. Sader, M. Merk, W. Ziegler, R. Busch, H. F. Zeilhofer, and H. H. Horch. "Sprechgeschwindigkeit bei kompensatorischer Artikulation von Patienten mit Lippen-Kiefer-Gaumenspalten". *Folia Phoniatrica et Logopaedica*, Vol. 51, pp. 272–286, 1999.

[Brunn 05] M. Brunner, A. Stellzig-Eisenhauer, U. Pröschel, R. Verres, and G. Komposch. "The Effect of Nasopharyngoscopic Biofeedback in Patients With Cleft Palate and Velopharyngeal Dysfunction". *Cleft Palate-Craniofacial Journal*, Vol. 42, No. 6, pp. 649–657, 2005.

[Bucha 84] B. Buchanan and E. Shortliffe. *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley Publishing Company, Reading, Massachusetts, Menlo Park, California, London, Amsterdam, Don Mills, Ontario, Sydney, 1984.

[Buder 00] E. Buder. "Acoustic Analysis of Voice Quality: A Tabulation of Algorithms 1902–1990". In: R. Kent and M. Ball, Eds., *Voice Quality Measurement*, Chap. 9, pp. 119–244, Singular Publishing Group, San Diego, USA, 2000.

[Burge 98] C. Burges. "A Tutorial on Support Vector Machines for Pattern Recognition". *Data Mining and Knowledge Discovery*, Vol. 2, No. 2, pp. 121–167, 1998.

[Buzo 80] A. Buzo, A. Gray, R. Gray, and J. Markel. "Speech Coding Based upon Vector Quantization". *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 28, No. 5, pp. 562–574, 1980.

[Cairn 94] D. Cairns and J. Hansen. "Nonlinear analysis and classification of speech under stressed conditions". *Journal of the Acoustic Society of America*, Vol. 96, No. 6, pp. 3392–3400, 1994.

[Cairn 96a] D. Cairns, J. Hansen, and J. Kaiser. “Recent Advances in Hypernasal Speech Detection using the Nonlinear Teager Energy Operator”. In: *Proceedings of the International Conference on Speech Communication and Technology (Interspeech)*, pp. 780–783, ISCA, Philadelphia, USA, 1996.

[Cairn 96b] D. Cairns, J. Hansen, and J. Riski. “A Noninvasive Technique for Detecting Hypernasal Speech using a nonlinear Operator”. *IEEE Transactions on Biomedical Engineering*, Vol. 43, No. 1, pp. 35–45, 1996.

[Canep 05] L. Canepari. *A Handbook of Pronunciation.* LINCOM GmbH, Muenchen, Germany, 1st Ed., 2005.

[Carin 03] F. Carinci, F. Pezzetti, and L. Scapoli. “Recent development in orofacial cleft genetics”. *J Craniofac Surg*, Vol. 14, No. 130, 2003.

[Catte 66] B. Cattel. “The scree test for the number of factors”. *Multivariate Behaviorial Research*, Vol. 1, pp. 245–276, 1966.

[Catts 97] H. W. Catts. “The Early Identification of Language-Based Reading Disabilities”. *Language, Speech, and Hearing Services in Schools*, Vol. 28, No. 1, pp. 86–89, 1997.

[Chen 05] Z.-H. Chen, Y.-F. Liao, and Y.-T. Juang. “Prosody Modeling and Eigen-Prosody Analysis for Robust Speaker Recognition”. In: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 185–188, IEEE Computer Society Press, Philadelphia, USA, 2005.

[Chen 06] Z.-H. Chen, Z.-R. Zeng, Y.-F. Liao, and Y.-T. Juang. “Probabilistic Latent Prosody Analysis for Robust Speaker Verification”. In: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 105–108, IEEE Computer Society Press, Toulouse, France, 2006.

[Cicch 76] D. Cicchetti. “Assessing inter-rater reliability for rating scales: Resolving some basic issues”. *British Journal of Psychiatry*, Vol. 129, No. 5, pp. 452–456, 1976.

[Cinca 02] T. Cincarek. “Klassifikation von Sprechergruppen”. Studienarbeit, Chair of Pattern Recognition, University Erlangen–Nuremberg, Erlangen, Germany, 2002.

[Cinca 04a] T. Cincarek. “Pronunciation Scoring for Non-Native Speech”. Diplomarbeit, Chair of Pattern Recognition, University Erlangen–Nuremberg, Erlangen, Germany, 2004.

[Cinca 04b] T. Cincarek, R. Gruhn, C. Hacker, E. Nöth, and S. Nakamura. “Gaikokugohatsuon no jidouhyoutei to yomiayamatta tango no jidoukenshutsu”. In: *Proceedings of the Acoustical Society of Japan*, pp. 165–166, September 2004.

[Clark 04] V. Clark, Ed. *SAS/STAT®9.1 User's Guide.* SAS Institute, Cary, NC, USA, 2004.

[Cohen 60] J. Cohen. “A Coefficient of Agreement for Nominal Scales”. *Educational and Psychological Measurement*, Vol. 20, No. 1, pp. 37–46, 1960.

[Cohen 83] J. Cohen and P. Cohen. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences.* Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1983.

[Courr 05] P. Courrieu. "Fast Computation of Moore Penrose Inverse Matrices". *Neural Information Processing*, Vol. 8, No. 2, pp. 25–29, 2005.

[Davie 82] M. Davies and J. Fleiss. "Measuring agreement for multinomial data". *Biometrics*, Vol. 38, No. 4, pp. 1047–1051, 1982.

[Davis 80] S. Davis and P. Mermelstein. "Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences". *IEEE Trans. on Acoustics, Speech, and Signal Processing (ASSP)*, Vol. 28, No. 4, pp. 357–366, 1980.

[Demps 77] A. Dempster, N. Laird, and D. Rubin. "Maximum Likelihood from Incomplete Data via the EM Algorithm". *J. Royal Statist. Soc. Ser. B*, Vol. 39, No. 1, pp. 1–22, 1977.

[Deng 05] J. Deng, T. Zheng, Z. Song, and J. Liu. "Modeling High-Level Information by using Gaussian Mixture Correlation for GMM-UBM based Speaker Recognition". In: *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pp. 2033–2036, ISCA, Lisbon, Portugal, 2005.

[Dolli 02] M. Döllinger, U. Hoppe, F. Hettlich, J. Lohscheller, S. Schuberth, and U. Eysholdt. "Vibration parameter extraction from endoscopic image series of the vocal folds". *IEEE Trans. on Biomedical Engineering*, Vol. 49, No. 8, pp. 773–781, 2002.

[Dolli 08] M. Döllinger, F. Rosanowski, U. Eysholdt, and J. Lohscheller. "Basic research on vocal fold dynamics: Three-dimensional vibration analysis of human and canine larynges". *HNO*, Vol. 56, No. 12, pp. 1213–1220, 2008.

[Eide 96] E. Eide and H. Gish. "A Parametric Approach to Vocal Tract Length Normalization". In: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 346–348, IEEE Computer Society Press, Atlanta, Georgia, USA, 1996.

[Epple 05] B. Eppley, J. van Aalst, A. Robey, R. Havlik, and M. Sadove. "The Spectrum of orofacial Clefting". *Plastic and Reconstructive Surgery*, Vol. 115, No. 7, pp. 101–114, 2005.

[Exner 07] J. Exner. "Visualization of Voice and Speech Disorders". Studienarbeit, Chair of Pattern Recognition, University Erlangen–Nuremberg, Erlangen, Germany, 2007.

[Eysho 08] U. Eysholdt and J. Lohscheller. "Phonovibrogram: vocal fold dynamics integrated within a single image". *HNO*, Vol. 56, No. 12, pp. 1207–1212, 2008.

[Fant 60a] G. Fant. *Acoustic Theory of Speech Production.* Mouton, The Hague, The Netherlands, 1960.

[Fant 60b] G. Fant. "Nasal Sounds and Nasalization". In: *Acoustic Theory of Speech Production*, Mouton, The Hague, The Netherlands, 1960.

[Fant 73] G. Fant. *Speech Sounds and Features.* The MIT Press, Cambridge (Massachusetts), London, 1973.

[Fawce 06] A. Fawcett. "An introduction to ROC analysis". *Pattern Recognition Letters*, Vol. 27, pp. 861–874, 2006.

[Fisch 05] M. Fischera, U. Hoppe, U. Eysholdt, and F. Rosanowski. "Tactile-Kinesthetic Responsiveness in Children with Cleft Lip Palate". *Laryng-Rhino-Otol*, Vol. 84, pp. 239–245, 2005.

[Fiscu 97] J. Fiscus. "A Post-Processing System to Yield Reduced Word Error Rates: Recogniser Output Voting Error Reduction". In: *Proc. IEEE ASRU Workshop*, pp. 347–352, Santa Barbara, USA, 1997.

[Fleis 69] J. Fleiss, J. Cohen, and B. Everitt. "Large sample standard errors of kappa and weighted kappa". *Psychological Bulletin*, Vol. 72, No. 5, pp. 323–327, 1969.

[Fleis 71] J. Fleiss. "Measuring Nominal Scale Agreement Among Many Raters". *Psychological Bulletin*, Vol. 76, No. 5, pp. 378–382, 1971.

[Fletc 87] R. Fletcher. *Practical Methods of Optimization.* John Wiley and Sons, Inc., Chichester, West Sussex, United Kingdom, 2nd Ed., 1987.

[Fletc 89] S. Fletcher, L. Adams, and M. McCutcheon. "Cleft Palate Speech Assessment Through Oral-Nasal Acoustic Measures". In: *Communicative Disorders Related to Cleft Lip and Palate*, pp. 246–257, Little and Brown, Boston, USA, 1989.

[Fox 02] A. Fox. "PLAKSS – Psycholinguistische Analyse kindlicher Sprechstörungen". Swets & Zeitlinger, Frankfurt a.M., Germany, now available from Harcourt Test Services GmbH, Germany, 2002.

[Frank 98a] E. Frank, Y. Wang, S. Inglis, G. Holmes, and I. Witten. "Using model trees for classification". *Machine Learning*, Vol. 32, No. 1, pp. 63–76, 1998.

[Frank 98b] E. Frank and I. H. Witten. "Generating Accurate Rule Sets Without Global Optimization". In: J. Shavlik, Ed., *Fifteenth International Conference on Machine Learning*, pp. 144–151, Morgan Kaufmann, 1998.

[Freun 96] Y. Freund and R. E. Schapire. "Experiments with a new boosting algorithm". In: *Thirteenth International Conference on Machine Learning*, pp. 148–156, Morgan Kaufmann, San Francisco, 1996.

[Fukun 90] K. Fukunaga. *Introduction to Statistical Pattern Recognition.* Academic Press, Boston, USA, 1990.

[Furui 00] S. Furui. *Digital Speech Processing.* Marcel Dekker, New York, USA, 2nd Ed., 2000.

[Furui 05] S. Furui. "50 years of progress in speech and speaker recognition". In: *Proceedings of the 10th International Conference on Speech and Computer (SPECOM)*, pp. 1–9, 2005.

[Furui 91] S. Furui. "Speaker-Independent and Speaker-Adaptive Recognition Techniques". In: S. Furui and M. M. Sondhi, Eds., *Advances in Speech Signal Processing*, pp. 597–622, Marcel Dekker, New York, USA, 1991.

[Furui 94] S. Furui. “An Overview of Speaker Recognition Technology”. In: *ESCA Workshop on Automatic Speaker Recognition, Identification, and Verification*, pp. 1–10, ESCA, Martigny, Switzerland, 1994.

[Furui 97] S. Furui. “Recent Advances in Speaker Recognition”. In: *Proc. First Int. Conf. Audio- and Video-based Biometric Person Authentication*, pp. 237–252, Crans-Montana, Switzerland, 1997.

[Gales 96] M. Gales, D. Pye, and P. Woodland. “Variance Compensation within the MLLR Framework for Robust Speech Recognition and Speaker Adaptation”. In: *Proceedings of the International Conference on Speech Communication and Technology (Interspeech)*, pp. 1832–1835, ISCA, Philadelphia, USA, 1996.

[Gales 97] M. Gales. “Maximum Likelihood Linear Transformations for HMM-based Speech Recognition”. Tech. Rep., Cambridge University Engineering Dept., Cambridge, UK, 1997.

[Gallw 02] F. Gallwitz. *Integrated Stochastic Models for Spontaneous Speech Recognition.* Vol. 6 of *Studien zur Mustererkennung*, Logos Verlag, Berlin, Germany, 2002.

[Gauva 94] J. Gauvain and C. 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.

[Ghise 64] E. Ghiselli. *Theory of Psychological Measurement.* McGraw-Hill Book Company, New York, USA, 1964.

[Gibbo 02] F. E. Gibbon and L. Crampin. “Labial-Lingual Double Articulations in Speakers with Cleft Palate”. *Cleft Palate-Craniofacial Journal*, Vol. 39, No. 1, pp. 40–49, 2002.

[Gray 18] H. Gray. *Henry Gray’s Anatomy of the Human Body.* Lea & Febiger, Philadelphia, PA, USA, 20th Ed., 1918.

[Haapa 96] M.-L. Haapanen, L. Liu, T. Hiltunen, L. Leinonen, and J. Karhunen. “Cul-de-sac hypernasality test with pattern recognition of LPC indices”. *Folia Phoniatr Logop*, Vol. 48, No. 1, pp. 35–43, 1996.

[Haas 07] J. Haas. “Experiences with Automatic Dialogue Systems: Theory and Practice”. 2007. Presentation at the Chair of Pattern Recognition.

[Hacke 05a] C. Hacker, A. Batliner, S. Steidl, E. Nöth, H. Niemann, and T. Cincarek. “Assessment of Non-Native Children’s Pronunciation: Human Marking and Automatic Scoring”. In: *Proceedings of the 10th International Conference on Speech and Computer (SPECOM)*, pp. 61–64, 2005.

[Hacke 05b] C. Hacker, T. Cincarek, R. Gruhn, S. Steidl, E. Nöth, and H. Niemann. “Pronunciation Feature Extraction”. In: G. Kropatsch, R. Sablatnig, and A. Hanbury, Eds., *Pattern Recognition, 27th DAGM Symposium, August 30 - September 2005, Vienna, Austria, Proceedings*, pp. 141–148, Springer, Berlin, Heidelberg,Germany, 2005.

[Hacke 06] C. Hacker, A. Batliner, and E. Nöth. “Are You Looking at Me, are You Talking with Me – Multimodal Classification of the Focus of Attention”. In: P. Sojka, I. Kopeček, and K. Pala, Eds., *9th International Conf. on Text, Speech and Dialogue (TSD)*, pp. 581 – 588, Springer, Berlin, Heidelberg, New York, 2006.

[Hacke 07a] C. Hacker, T. Cincarek, A. Maier, A. Heßler, and E. Nöth. “Boosting of Prosodic and Pronunciation Features to Detect Mispronunciations of Non-Native Children”. In: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 197–200, IEEE Computer Society Press, Hawaii, USA, 2007.

[Hacke 07b] C. Hacker, A. Maier, A. Hessler, U. Guthunz, and E. Nöth. “Caller: Computer Assisted Language Learning from Erlangen - Pronunciation Training and More”. In: *International Conference on Interactive Computer Aided Learning*, Kassel University Press, Kassel, Germany, 2007. no pagination.

[Hader 02] T. Haderlein. “Using the ISADORA System for Analyzing Fatigue Symptoms and Robustness of Features against Reverberation”. Tech. Rep., Chair of Multimedia Communications and Signal Processing, University Erlangen-Nuremberg, 2002.

[Hader 04] T. Haderlein, S. Steidl, E. Nöth, F. Rosanowski, and M. Schuster. “Automatic Recognition and Evaluation of Tracheoesophageal Speech”. In: P. Sojka, I. Kopeček, and K. Pala, Eds., *7th International Conf. on Text, Speech and Dialogue (TSD)*, pp. 331–338, Springer, Berlin, Heidelberg, New York, 2004.

[Hader 06a] T. Haderlein, E. Nöth, M. Schuster, U. Eysholdt, and F. Rosanowski. “Evaluation of Tracheoesophageal Substitute Voices Using Prosodic Features”. In: R. Hoffmann and H. Mixdorff, Eds., *Proc. Speech Prosody, 3rd International Conference*, pp. 701–704, TUDpress, Dresden, Germany, 2006.

[Hader 06b] T. Haderlein, D. Zorn, S. Steidl, E. Nöth, M. Shozakai, and M. Schuster. “Visualization of Voice Disorders Using the Sammon Transform”. In: P. Sojka, I. Kopeček, and K. Pala, Eds., *9th International Conf. on Text, Speech and Dialogue (TSD)*, pp. 589–596, Springer, Berlin, Heidelberg, New York, 2006.

[Hader 07a] T. Haderlein. *Automatic Evaluation of Tracheoesophageal Substitute Voices.* Vol. 25 of *Studien zur Mustererkennung*, Logos Verlag, Berlin, Germany, 2007.

[Hader 07b] T. Haderlein, K. Riedhammer, A. Maier, E. Nöth, H. Toy, and F. Rosanowski. “An Automatic Version of the Post-Laryngectomy Telephone Test”. In: V. Matoušek and P. Mautner, Eds., *10th International Conf. on Text, Speech and Dialogue (TSD)*, pp. 238–245, Springer, Berlin, Heidelberg, New York, 2007.

[Hader 08] T. Haderlein, E. Nöth, A. Maier, S. Schuster, and F. Rosanowski. “Influence of Reading Errors on the Text-Based Automatic Evaluation of Pathologic Voices”. In: P. Sojka, I. Kopeček, and K. Pala, Eds., *11th International Conf. on Text, Speech and Dialogue (TSD)*, pp. 325–332, Springer, Berlin, Heidelberg, New York, 2008.

[Halbe 04] B. Halberstam. “Acoustic and Perceptual Parameters Relating to Connected Speech Are More Reliable Measures of Hoarseness than Parameters Relating to Sustained Vowels”. *ORL - Journal for Oto-Rhino-Laryngology and Its Related Specialties*, Vol. 66, No. 2, pp. 70–73, 2004.

[Hall 98] M. A. Hall. *Correlation-based Feature Subset Selection for Machine Learning.* PhD thesis, University of Waikato, Hamilton, New Zealand, 1998.

[Hardi 92] M. Hardin, D.-R. Van Demark, H. Morris, and M. Payne. "Correspondence between Nasalance Scores and Listener Judgments of Hypernasality and Hyponasality". *Cleft Palate-Craniofacial Journal*, Vol. 29, No. 4, pp. 346–351, 1992.

[Hardi 98] A. Harding and P. Grunwell. "Active versus passive cleft-type speech characteristics". *Int J Lang Commun Disord*, Vol. 33, No. 3, pp. 329–352, 1998.

[Hausa 00] J. Hausamen. *Lippen-Kiefer-Gaumenspalten – Informationsbroschüre über die Behandlung in der Medizinischen Hochschule Hannover*. Medizinische Hochschule Hannover. Hannover, Germany, 2000.

[Hertl 99] H. Hertlein. "Textunabhängige Sprecheridentifikation zur Zugangskontrolle". Diplomarbeit, Chair of Pattern Recognition, University Erlangen–Nuremberg, Erlangen, Germany, 1999.

[Hessl 05] A. Heßler. "Entwicklung einer Englisch-Lernsoftware mit integrierter Spracherkennung". Studienarbeit, Chair of Pattern Recognition, University Erlangen–Nuremberg, Erlangen, Germany, 2005.

[Hiran 81] M. Hirano. *Clinical Examination of Voice*. Springer, New York, NY, USA, 1981.

[Hogen 04] T. Hogen Esch and P. Dejonckere. "Objectivating Nasality in Healthy and Velopharyngeal Insufficient Children with the Nasalance Acquisition System (NasalView): Defining Minimal Required Speech Tasks Assessing Normative Values for Dutch Language". *Int J Pediatr Otorhinolaryngol*, Vol. 68, No. 8, pp. 1039–1046, 2004.

[Holte 93] R. Holte. "Very simple classification rules perform well on most commonly used datasets". *Machine Learning*, Vol. 11, pp. 63–91, 1993.

[Horii 81] Y. Horii and J. Lang. "Distributional Analysis of an Index of Nasal Coupling (HONC) in Simulated Hypernasal Speech". *Cleft Palate J*, Vol. 18, No. 4, pp. 279–285, 1981.

[Horii 83] Y. Horii. "An Accelerometric Measure as a Physical Correlate of Perceived Hypernasality in Speech". *Journal of Speech and Hearing Research*, Vol. 26, pp. 476–480, 1983.

[Huang 01] X. Huang, A. Acero, and H.-W. Hon. *Spoken Language Processing – A Guide to Theory, Algorithm, and System Development*. Prentice Hall, Upper Saddle River, 2001.

[Huber 02] R. Huber. *Prosodisch-linguistische Klassifikation von Emotion*. Vol. 8 of *Studien zur Mustererkennung*, Logos Verlag, Berlin, Germany, 2002.

[Imato 99] S. Imatomi, T. Arai, Y. Mimura, and M. Kato. "Effects of Hoarseness on Hypernasality Ratings". In: *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pp. 1075–1078, ISCA, Budapest, Hungary, 1999.

[ITU 01] ITU. "PESQ, An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs". P.862, 2001.

[ITU 04] ITU. "Single-ended method for objective speech quality narrow-band telephony applications". P.563, 2004.

[Jacob 97] B. Jacobson, A. Johnson, C. Grywalski, A. Silbergleit, G. Jacobson, M. Benninger, and C. Newman. "The Voice Handicap Index (VHI): Development and Validation". *Am J Speech-Language Path*, Vol. 6, No. 3, pp. 66–70, 1997.

[John 95] G. H. John and P. Langley. "Estimating Continuous Distributions in Bayesian Classifiers". In: *Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 338–345, Morgan Kaufmann, San Mateo, 1995.

[Karli 93a] J. Karling, O. Larsen, R. Leanderson, K. Galyas, and A. Serpa-Leitao. "NORAM—An Instrument Used in the Assessment of Hypernasality". *Cleft Palate J*, Vol. 30, No. 2, 1993.

[Karli 93b] J. Karling, O. Larson, R. Leanderson, and G. Henningsson. "Speech in Unilateral and Bilateral Cleft Palate Patients from Stockholm". *Cleft Palate-Craniofacial Journal*, Vol. 30, No. 1, pp. 73–77, 1993.

[Karne 05] M. Karnell, P. Bailey, L. Johnson, A. Dragan, and J. Canady. "Facilitating Communication Among Speech Pathologists Treating Children With Cleft Palate". *Cleft Palate-Craniofacial Journal*, Vol. 42, No. 6, pp. 585–588, 2005.

[Karne 95] M. Karnell. "Nasometric Discrimination of Hypernasality and Turbulent Nasal Airflow". *Cleft Palate-Craniofacial Journal*, Vol. 32, No. 2, pp. 145–148, 1995.

[Katao 96] R. Kataoka, K. Michi, K. Okabe, T. Miura, and H. Yoshida. "Spectral Properties and Quantitative Evaluation of Hypernasality in Vowels". *Cleft Palate-Craniofacial Journal*, Vol. 33, No. 1, pp. 43–50, 1996.

[Kawam 90] H. Kawamoto. "Rare craniofacial clefts". In: J. C. McCarthy, Ed., *Plastic Surgery*, Saunders, Philadelphia, USA, 1990.

[Kay 93] *Multi-Dimensional Voice Program (MDVP), Model 4305 [Computer software]*. Kay Elemetrics Corporation, New York, USA, 1993.

[Kay 94] *Instruction manual of the nasometer Model 6200-3, IBM PC Version*. Kay Elemetrics Corporation, New York, USA, 1994.

[Kiess 97] A. Kießling. *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung. Berichte aus der Informatik*, Shaker, Aachen, Germany, 1997.

[Kitzi 09] P. Kitzing, A. Maier, and V. Ahlander. "Automatic Speech Recognition (ASR) and its Use as a Tool for Assessment or Therapy of Voice, Speech and Language Disorders". *Logopedics Phoniatrics Vocology*, 2009. to appear.

[Knese 93] R. Kneser and V. Steinbiss. "On the Dynamic Adaptation of Stochastic Language Models". In: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 586–588, IEEE Computer Society Press, Minneapolis, USA, 1993.

[Kripp 03] K. Krippendorff. *Content Analysis, an Introduction to its Methodology*. Sage Publications, Thousand Oaks, USA, 2nd Ed., 2003.

[Krumm 99] F. Krummenauer. "Erweiterungen von Cohen's kappa-Maß für Multi-Rater-Studien: Eine Übersicht". *Informatik, Biometrie und Epidemiologie in Medizin und Biologie*, Vol. 30, No. 1, pp. 3–20, 1999.

[Kuttn 03] C. Küttner, R. Schönweiler, B. Seeberger, R. Dempf, J. Lisson, and M. Ptok. "Objektive Messung der Nasalanz in der deutschen Hochlautung". *HNO*, Vol. 51, pp. 151–156, 2003.

[Laiti 00] J. Laitinen, R. Ranta, J. Pulkkinen, M. Paaso, and M. Haapanen. "Changes in Finnish Dental Consonant Articulation in Cleft Lip/Palate Children between 6 and 8 Years of Age". *Folia Phoniatrica et Logopaedica*, Vol. 52, pp. 253–259, 2000.

[Laiti 06] J. Laitinen, R. Ranta, J. Pulkkinen, M. Paaso, and M. Haapanen. "Conversational Skills of Children with Cleft Lip and Palate: A Replication and Extension". *Cleft Palate-Craniofacial Journal*, Vol. 43, pp. 179–188, 2006.

[Laiti 98] J. Laitinen, M. Haapanen, M. Paaso, J. Pulkkinen, A. Heliövaara, and R. Ranta. "Occurrence of Dental Consonant Misarticulations in Different Cleft Types". *Folia Phoniatrica et Logopaedica*, Vol. 50, pp. 92–100, 1998.

[Lee 97] S. Lee, A. Potamianos, and S. Narayanan. "Analysis of Children's Speech: Duration, Pitch and Formants". In: *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pp. 473–476, ISCA, Rhodes, Greece, 1997.

[Levit 00] M. Levit. "Benutzung von Sprachcharakteristika zur Klassifikation von Sprachvarietäten und Sprecherzuständen". Diplomarbeit, Chair of Pattern Recognition, University Erlangen–Nuremberg, Erlangen, Germany, 2000.

[Li 01] Q. Li and M. Russell. "Why is Automatic Speech Recognition of Children's Speech Difficult?". In: *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pp. 2671–2674, ISCA, Aalborg, Denmark, 2001.

[Lierd 01] K. V. Lierde, J. V. Borsel, P. V. Cauwenberge, and S. Callewaert. "Speech Patterns in Children with Velo-Cardio-Facial Syndrome". *Folia Phoniatrica et Logopaedica*, Vol. 53, pp. 213–221, 2001.

[Lierd 02] K. V. Lierde, M. D. Bodt, J. V. Borsel, F. Wuyts, and P. V. Cauwenberge. "Effect of cleft type on overall speech intelligibility and resonance". *Folia Phoniatrica et Logopaedica*, Vol. 54, No. 3, pp. 158–168, 2002.

[Lierd 03] K. V. Lierde, M. D. Bodt, I. Baetens, V. Schrauwen, and P. V. Cauwenberge. "Outcome of Treatment regarding Articulation, Resonance and Voice in Flemish Adults with Unilateral and Bilateral Cleft Palate". *Folia Phoniatrica et Logopaedica*, Vol. 55, pp. 80–90, 2003.

[Lierd 04] K. V. Lierde, S. Monstrey, K. Bonte, P. V. Cauwenberge, and B. Vinck. "The long-term speech outcome in Flemish young adults after two different types of palatoplasty". *Int J Pediatr Otorhinolaryngol*, Vol. 68, pp. 865–875, 2004.

[Liker 32] R. Likert. "A technique for the measurement of attitudes". *Archives of Psychology*, Vol. 140, 1932. Columbia University, New York, NY, USA.

[Linde 80] Y. Linde, A. Buzo, and R. Gray. "An Algorithm for Vector Quantizer Design". *IEEE Trans. on Communications*, Vol. 28, No. 1, pp. 84–95, 1980.

[Liu 96] H. Liu and R. Setiono. "A probabilistic approach to feature selection - A filter solution". In: *13th International Conference on Machine Learning*, pp. 319–327, 1996.

[Lohma 02] A. Lohmander, C. Persson, and P. Owman-Moll. "Unrepaired clefts in the hard palate: speech deficits at the age of 5 and 7 years and their relationship to the size of the cleft". *Scand J Plast Reconstr Surg Hand Surg*, Vol. 36, pp. 332–339, 2002.

[Lohsc 09] J. Lohscheller and U. Eysholdt. "Phonovibrogram Visualization of Entire Vocal Fold Dynamics". *Laryngoscope*, Vol. 4, No. 118, pp. 753–759, 2009.

[Lu 04] Y. Lu. "Recognition of nasality and nasal air flow problems in children with cleft lip and palate and/or velopharngeal incompetence". Master Thesis, University of Sheffield, United Kingdom, 2004.

[Maier 05a] A. Maier. "Recognizer Adaptation by Acoustic Model Interpolation on a Small Training Set from the Target Domain". Diplomarbeit, Chair of Pattern Recognition, University Erlangen–Nuremberg, Erlangen, Germany, 2005.

[Maier 05b] A. Maier. "Robust Speech Recognition of Noisy or Reverberated Data Using Multiple Recognizers in Different Energy Bands". Studienarbeit, Chair of Pattern Recognition, University Erlangen–Nuremberg, Erlangen, Germany, 2005.

[Maier 05c] A. Maier, C. Hacker, S. Steidl, and E. Nöth. "Helfen "Fallen" bei verrauschten Daten? — Spracherkennung mit TRAPs". In: *Fortschritte der Akustik — Proc. DAGA '05*, pp. 315–316, Munich, Germany, 2005.

[Maier 05d] A. Maier, C. Hacker, S. Steidl, E. Nöth, and H. Niemann. "Robust Parallel Speech Recognition in Multiple Energy Bands". In: G. Kropatsch, R. Sablatnig, and A. Hanbury, Eds., *Pattern Recognition, 27th DAGM Symposium, August 30 - September 2005, Vienna, Austria, Proceedings*, pp. 133–140, Springer, Berlin, Heidelberg,Germany, 2005.

[Maier 06a] A. Maier. "PEAKS - Programm zur Evaluation und Analyse Kindlicher Sprachstörungen - Bedienungsanleitung". Tech. Rep. 1, University of Erlangen-Nuremberg, Erlangen, 2006.

[Maier 06b] A. Maier, C. Hacker, E. Nöth, E. Nkenke, T. Haderlein, F. Rosanowski, and M. Schuster. "Intelligibility of Children with Cleft Lip and Palate: Evaluation by Speech Recognition Techniques". In: *Proc. International Conf. on Pattern Recognition (ICPR)*, pp. 274–277, Hong Kong, China, 2006.

[Maier 06c] A. Maier, T. Haderlein, C. Hacker, E. Nöth, F. Rosanowski, U. Eysholdt, and M. Schuster. "Automatische internetbasierte Evaluation der Verständlichkeit". In: M. Gross and F. Kruse, Eds., *Aktuelle phoniatrisch-pädaudiologische Aspekte 2006*, pp. 87–90, Books On Demand GmbH, Norderstedt, Germany, 2006.

[Maier 06d] A. Maier, T. Haderlein, and E. Nöth. "Environmental Adaptation with a Small Data Set of the Target Domain". In: P. Sojka, I. Kopeček, and K. Pala, Eds., *9th International Conf. on Text, Speech and Dialogue (TSD)*, pp. 431–437, Springer, Berlin, Heidelberg, New York, 2006.

[Maier 06e] A. Maier, E. Nöth, A. Batliner, E. Nkenke, and M. Schuster. "Fully Automatic Assessment of Speech of Children with Cleft Lip and Palate". *Informatica*, Vol. 30, No. 4, pp. 477–482, 2006.

[Maier 06f] A. Maier, E. Nöth, E. Nkenke, and M. Schuster. "Automatic Assessment of Children's Speech with Cleft Lip and Palate". In: T. Erjavec and J. Žganec Gros, Eds., *Information Society Language Technologies Conference (IS-LTC)*, pp. 31–35, Ljubljana, Slovenia, 2006.

[Maier 07a] A. Maier, T. Haderlein, M. Schuster, E. Nkenke, and E. Nöth. "Intelligibility is more than a single Word: Quantification of Speech Intelligibility by ASR and Prosody". In: V. Matoušek and P. Mautner, Eds., *10th International Conf. on Text, Speech and Dialogue (TSD)*, pp. 278–285, Springer, Berlin, Heidelberg, New York, 2007.

[Maier 07b] A. Maier, T. Haderlein, M. Schuster, and E. Nöth. "PEAKS—A Platform for Evaluation and Analysis of all Kinds of Speech Disorders". In: *Proc. 41st Annual Meeting of the Society for Biomedical Technologies of the Association for Electrical, Electronic & Information Technologies (BMT 2007)*, Aachen, Germany, 2007. no pagination.

[Maier 07c] A. Maier, E. Nöth, U. Eysholdt, and M. Schuster. "Automatische Bewertung der Nasalität von Kindersprache". In: M. Gross and F. Kruse, Eds., *Aktuelle phoniatrisch-pädaudiologische Aspekte 2007*, pp. 74–76, Books On Demand GmbH, Norderstedt, Germany, 2007.

[Maier 07d] A. Maier, M. Schuster, A. Batliner, E. Nöth, and E. Nkenke. "Automatic Scoring of the Intelligibility in Patients with Cancer of the Oral Cavity". In: *Interspeech 2007 – Proc. Int. Conf. on Spoken Language Processing, 10th European Conference on Spoken Language Processing, August 27-31, 2007, Antwerp, Belgium, Proceedings*, pp. 1206–1209, 2007.

[Maier 08a] A. Maier. *Parallel Robust Speech Recognition.* VDM Verlag Dr. Müller, Saarbrücken, Germany, 2008.

[Maier 08b] A. Maier. *Speech Recognizer Adaptation.* VDM Verlag Dr. Müller, Saarbrücken, Germany, 2008.

[Maier 08c] A. Maier, J. Exner, S. Steidl, A. Batliner, T. Haderlein, and E. Nöth. "An Extension to the Sammon Mapping for the Robust Visualization of Speaker Dependencies". In: P. Sojka, I. Kopeček, and K. Pala, Eds., *11th International Conf. on Text, Speech and Dialogue (TSD)*, pp. 381–388, Springer, Berlin, Heidelberg, New York, 2008. to appear.

[Maier 08d] A. Maier, T. Haderlein, E. Nöth, F. Rosanowski, U. Eysholdt, and M. Schuster. "Visualisierung der Differenzierung von Stimm- und Sprechbefunden". In: M. Gross and F. Kruse, Eds., *Aktuelle phoniatrisch-pädaudiologische Aspekte 2008*, pp. 191–192, Rheinware Verlag, Mönchengladbach, Germany, 2008.

[Maier 08e] A. Maier, T. Haderlein, E. Nöth, and M. Schuster. "PEAKS: Ein Client-Server-Internetportal zur Berwertung der Aussprache". In: *Telemed 2008, Proceedings*, pp. 104–107, Akademische Verlagsgesellschaft, Aka GmbH, Heidelberg, Germany, 2008.

[Maier 08f] A. Maier, T. Haderlein, F. Rosanowski, C. Sous-Kulke, and W. Schupp. "Automatische Bewertung der Aussprache von Patienten mit Dysarthrie". In: *8. Jahrestagung der Gesellschaft für*

Aphasieforschung und -behandlung, Nuremberg, Germany, 2008. to appear.

[Maier 08g] A. Maier, F. Hönig, C. Hacker, M. Schuster, and E. Nöth. "Automatic Evaluation of Characteristic Speech Disorders in Children with Cleft Lip and Palate". In: *Interspeech 2008 – Proc. Int. Conf. on Spoken Language Processing, 11th International Conference on Spoken Language Processing, September 25-28, 2008, Brisbane, Australia, Proceedings*, pp. 1757–1760, 2008.

[Maier 08h] A. Maier, A. Reuss, C. Hacker, M. Schuster, and E. Nöth. "Analysis of Hypernasal Speech in Children with Cleft Lip and Palate". In: P. Sojka, I. Kopeček, and K. Pala, Eds., *11th International Conf. on Text, Speech and Dialogue (TSD)*, pp. 389–396, Springer, Berlin, Heidelberg, New York, 2008.

[Maier 08i] A. Maier, M. Schuster, and E. Nöth. "Towards Monitoring of Children's Speech - A Case Study". In: *Proceedings of the 1st Workshop on Child, Computer and Interaction*, Chania, Crete, Greece, 2008. to appear.

[Milla 01] T. Millard and L. Richman. "Different cleft conditions, facial appearance, and speech: relationship to psychological variables". *Cleft Palate Craniofac J*, Vol. 38, pp. 68–75, 2001.

[Mille 02] A. Miller. *Subset selection in Regression.* Chapman & Hall / CRC, Boca Raton, London, New York, Washington, D.C., 2nd Ed., 2002.

[Moore 20] E. H. Moore. "On the reciprocal of the general algebraic matrix". *Bulletin of the American Mathematical Society*, Vol. 26, pp. 394–395, 1920.

[Moore 86] B. Moore. *Frequency Selectivity in Hearing.* Academic, London, England, 1986.

[Morri 03] H. Morris and A. Ozanne. "Phonetic, Phonological, and Language Skills of Children with a cleft palate". *Cleft Palate-Craniofacial Journal*, Vol. 40, No. 5, pp. 460–470, 2003.

[Nadas 85] A. Nadas. "On Turing's Formula for Word Probabilities". *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 33, No. 6, pp. 1414–1416, 1985.

[Nagin 05] M. Nagino, G. Shozakai. "Building an Effective Corpus by Using Acoustic Space Visualization (COSMOS) Method". In: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 449–452, IEEE Computer Society Press, Philadelphia, USA, 2005.

[Nakaj 01] T. Nakajima, A. Mitsudome, and A. Yoshikawa. "Postoperative speech development based on cleft types in children with cleft palate". *Pediatrics International*, Vol. 43, pp. 666–672, 2001.

[Naylo 07] W. Naylor and B. Chapman. "WNLIB Homepage". 2007.
http://www.willnaylor.com/wnlib.html, last visited 07/20/2007.

[Niema 03] H. Niemann. *Klassifikation von Mustern.* available online, 2nd Ed., 2003.
http://www5.informatik.uni-erlangen.de/Personen/niemann/klassifikation-von-mustern/m00links.html; last visited 02/12/2008.

[Niema 79] H. Niemann and J. Weiss. “A fast-converging algorithm for nonlinear mapping of highdimensional data to a plane”. *IEEE Trans. Computers*, Vol. C-28, pp. 142–147, 1979.

[Niema 93] H. Niemann, E. Nöth, E. Schukat-Talamazzini, A. Kießling, R. Kompe, T. Kuhn, K. Ott, and S. Rieck. “Statistical Modeling of Segmental and Suprasegmental Information”. In: *Proc. NATO ASI Conference “New Advances and Trends in Speech Recognition and Coding”*, pp. 237–260, Bubion, Granada, Spain, 1993.

[Noth 00] E. Nöth, H. Niemann, T. Haderlein, M. Decher, U. Eysholdt, F. Rosanowski, and T. Wittenberg. “Automatic Stuttering Recognition using Hidden Markov Models”. In: *Proceedings of the International Conference on Speech Communication and Technology (Interspeech)*, pp. 65–68, ISCA, Beijing, China, 2000.

[Noth 07] E. Nöth, A. Maier, T. Haderlein, K. Riedhammer, F. Rosanowski, and M. Schuster. “Automatic Evaluation of Pathologic Speech — from Research to Routine Clinical Use”. In: V. Matoušek and P. Mautner, Eds., *10th International Conf. on Text, Speech and Dialogue (TSD)*, pp. 294–301, Springer, Berlin, Heidelberg, New York, 2007.

[Noth 91] E. Nöth. *Prosodische Information in der automatischen Spracherkennung — Berechnung und Anwendung*. Niemeyer, Tübingen, 1991.

[Ogus 07] E. Ogus, A. Yazici, and F. Gurbuz. “Evaluating the Significance Test when the Correlation Coefficient is Different from Zero in the Test of the Hypothesis”. *Communications in Statistic—Simulation and Computation*, Vol. 36, No. 4, pp. 847–854, 2007.

[OPTI 04] “3SQM$^{\mathrm{TM}}$—ADVANCED NON-INTRUSIVE VOICE QUALITY TESTING”. Tech. Rep., OPTICOM GmbH, Erlangen, Germany, 2004. http://www.opticom.de/download/3SQM-WP-290604.pdf.

[Paal 05] S. Paal, U. Reulbach, K. Strobel-Schwarthoff, E. Nkenke, and M. Schuster. “Beurteilung von Sprechauffälligkeiten bei Kindern mit Lippen-Kiefer-Gaumen-Spaltbildungen”. *J Orofac Orthop*, Vol. 66, No. 4, pp. 270–278, 2005.

[Pahn 01] J. Pahn, R. Dahl, and E. Pahn. “Beziehung zwischen Messung der stimmlichen Durchdringungsfähigkeit, Stimmstatus nach Pahn und ausgewählten Parametern des Stimmanalyseprogramms MDVP (Kay)”. *Folia Phoniatr Logop*, Vol. 53, No. 6, pp. 308–316, 2001.

[Palio 05] V. Paliobei, A. Psifidis, and D. Anagnostopoulos. “Hearing and speech assessment of cleft palate patients after palatal closure”. *Int J of Pediatr Otorhinilaryngol*, Vol. 69, pp. 1373–1381, 2005.

[Pampl 00] M. Pamplona, A. Ysunza, M. González, E. Ramírez, and C. Patiño. “Linguistic development in cleft palate patients with and without compensatory articulation disorder”. *Int J Pediatr Otorhinolaryngol*, Vol. 54, pp. 81–91, 2000.

[Pampl 05] C. Pamplona, A. Ysunza, C. Patiño, E. Ramírez, M. Drucker, and J.Mazón. “Speech summer camp for treating articulation disorders in cleft palate patients”. *Int J Pediatr Otorhinolaryngol*, Vol. 69, pp. 351–359, 2005.

[Pears 01] K. Pearson. “On Lines and Planes of Closest Fit to Systems of Points in Space”. *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science*, Vol. 2, No. 6, pp. 559–572, 1901.

[Pears 96] K. Pearson. “Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia”. *Philosophical Transactions of the Royal Society of London*, Vol. 187, pp. 253–318, 1896.

[Penro 55] R. Penrose. “A generalized inverse for matrices”. *Proceedings of the Cambridge Philosophical Society*, Vol. 51, pp. 406–413, 1955.

[Peter 95] S. J. Peterson-Falzone. “Speech Outcomes in Adolescents with Cleft Lip and Palate”. *Cleft Palate-Craniofacial Journal*, Vol. 32, No. 3, pp. 125–128, 1995.

[Press 92] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. *Numerical Recipes in C.* Cambridge University Press, Cambridge, MA, USA, 1992.

[Pruth 03] T. Pruthi and C. Y. Espy-Wilson. “Automatic Classification of Nasals and Semivowels”. In: *ICPhS 2003 – 15th International Congress of Phonetic Sciences, August 2003, Barcelona, Spain, Proceedings*, pp. 3061–3064, 2003.

[Pruth 04] T. Pruthi and C. Y. Espy-Wilson. “Acoustic parameters for automatic detection of nasal manner”. *Speech Communication*, Vol. 43, No. 3, pp. 225–239, 2004.

[Pruth 07a] T. Pruthi. *Analysis, Vocal-tract Modeling and Automatic Detection of Vowel Nasalization.* PhD thesis, Graduate School of the University of Maryland, Maryland, USA, 2007.

[Pruth 07b] T. Pruthi and C. Y. Espy-Wilson. “Acoustic Parameters for the Automatic Detection of Vowel Nasalization”. In: *Interspeech 2007 – Proc. Int. Conf. on Spoken Language Processing, 10th European Conference on Spoken Language Processing, August 27-31, 2007, Antwerp, Belgium, Proceedings*, pp. 1925–1928, 2007.

[Pruth 07c] T. Pruthi, C. Y. Espy-Wilson, and H. Brad. “Story, Simulation and analysis of nasalized vowels based on magnetic resonance imaging data”. *J. Acoust. Soc. Am.*, Vol. 121, No. 6, pp. 3858–3873, 2007.

[Pulkk 01] J. Pulkkinen, M. Haapanen, M. Paaso, J. Laitinen, and R. Ranta. “Velopharyngeal Function from the Age of Three to Eight Years in Cleft Palate Patients”. *Folia Phoniatrica et Logopaedica*, Vol. 53, pp. 93–98, 2001.

[Pulkk 02] J. Pulkkinen, R. Ranta, M. Haapanen, A. Heliövaara, and J. Laitinen. “Associations between Lateral Cephalometric Dimensions and Misarticulations of Finnish Dental Consonants in Cleft Lip/Palate Children”. *Folia Phoniatrica et Logopaedica*, Vol. 54, pp. 240–246, 2002.

[Quinl 93] R. Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[Rasp 06] O. Rasp, J. Lohscheller, M. Döllinger, U. Eysholdt, and U. Hoppe. “The pitch rise paradigm: a new task for real-time endoscopy of non-stationary phonation”. *Folia Phoniatr Logop*, Vol. 58, No. 3, pp. 175–185, 2006.

[Reden 85] M. Redenbaugh and A. Reich. "Correspndence Between NAVI and Listeners' Direct Magnitude Estimations of Hypernasality". *Journal of Speech and Hearing Research*, Vol. 28, pp. 273–281, 1985.

[Redne 84] R. A. Redner and H. F. Walker. "Mixture Densities, Maximum Likelihood and the EM Algorithm". *Society for Industrial and Applied Mathematics Review*, Vol. 26, No. 2, pp. 195–239, 1984.

[Resco 01] E. Rescorla. *SSL and TLS: designing and building secure systems.* Addison-Wesley, Boston, USA, 2001.

[Reuss 07] A. Reuß. "Analysis of Speech Disorders in Children with Cleft Lip and Palate on Phoneme and Word Level". Studienarbeit, Chair of Pattern Recognition, University Erlangen–Nuremberg, Erlangen, Germany, 2007.

[Reyno 00] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. "Speaker Verification using Adapted Gaussian Mixture Models". *Digital Signal Processing*, pp. 19–41, 2000.

[Reyno 95] D. A. Reynolds and R. C. Rose. "Robust Test-Independent Speaker Identification using Gaussian Mixture Speaker Models". *IEEE Transaction on Speech and Audio Processing*, Vol. 3, pp. 72–83, 1995.

[Riedh 06] K. Riedhammer, T. Haderlein, M. Schuster, F. Rosanowski, and E. Nöth. "Automatic Evaluation of Tracheoesophageal Telephone Speech". In: T. Erjavec and J. Žganec Gros, Eds., *Information Society Language Technologies Conference (IS-LTC)*, pp. 17–22, Ljubljana, Slovenia, 2006.

[Riedh 07a] K. Riedhammer. "An Automatic Intelligibility Test Based on the Post-Laryngectomy Telephone Test". Studienarbeit, Chair for Pattern Recognition (Informatik 5), University of Erlangen-Nuremberg, Germany, 2007.

[Riedh 07b] 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 Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 717–722, IEEE Computer Society Press, Kyoto, Japan, 2007.

[Rijsb 79] C. J. Rijsbergen. *Information Retrieval.* Butterworths, London, UK, 1979.

[Rives 92] R. Rivest. "The MD5 Message-Digest Algorithm". RFC 1321 (Informational), 1992.

[Robbi 86] J. Robbins, J. Christensen, and G. Kempster. "Characteristics of speech production after tracheoesophageal puncture: voice onset time and vowel duration". *J Speech Hear Res*, Vol. 29, No. 4, pp. 499–504, 1986.

[Rosan 02] F. Rosanowski and U. Eysholdt. "Phoniatric aspects in cleft lip patients". *Facial Plast Surg*, Vol. 18, No. 3, pp. 197–203, 2002.

[Ruben 00] R. Ruben. "Redefining the survival of the fittest: communication disorders in the 21st century". *Laryngoscope*, Vol. 110, No. 2, pp. 241–245, 2000.

[Rupp 06] C. Rupp. *Requirements-Engineering und -Management. Professionelle, iterative Anforderungsanalyse für die Praxis.* Hanser Fachbuchverlag, München, Germany, 2006.

[Sammo 69] J. Sammon. "A nonlinear mapping for data structure analysis". *IEEE Trans. Computers*, Vol. C-18, pp. 401–409, 1969.

[Sanko 83] D. Sankoff and J. Kruskal, Eds. *Time Warps, String Edits, and Makromolecules.* Addison–Wesley, 1983.

[Schia 92] N. Schiavetti. "Scaling procedures for the measurement of speech intelligibility". In: R. D. Kent, Ed., *Intelligibility in Speech Disorders: Theory, measurement and management*, pp. 11–34, John Benjamins, Philadelphia, USA, 1992.

[Schol 97] B. Schölkopf. *Support Vector Learning.* PhD thesis, Technische Universität Berlin, Germany, 1997.

[Schon 94] R. Schönweiler and B. Schönweiler. "Hörvermögen und Sprachleistungen bei 417 Kindern mit Spaltfehlbildungen". *HNO*, Vol. 42, No. 11, pp. 691–696, 1994.

[Schon 99] R. Schönweiler, J. Lisson, B. Schönweiler, A. Eckardt, M. Ptok, J. Trankmann, and J. Hausamen. "A retrospective study of hearing, speech and language function in children with clefts following palatoplasty and veloplasty procedures at 18-24 months of age". *Int J Pediatr Otorhinolaryngol*, Vol. 50, No. 3, pp. 205–217, 1999.

[Schuk 95] E. G. Schukat-Talamazzini. *Automatische Spracherkennung – Grundlagen, statistische Modelle und effiziente Algorithmen.* Vieweg, Braunschweig, Germany, 1995.

[Schul 07] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl. "Towards more Reality in the Recognition of Emotional Speech". In: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 941–944, IEEE Computer Society Press, Hawaii, USA, 2007.

[Schus 03] M. Schuster, P. Kummer, U. Eysholdt, and F. Rosanowski. "Soziale Orientierung der Eltern von Kindern mit Lippen-Kiefer-Gaumen-Spalten". *HNO*, Vol. 51, pp. 507–511, 2003.

[Schus 05] M. Schuster, E. Nöth, T. Haderlein, S. Steidl, A. Batliner, and F. Rosanowski. "Can You Understand Him? Let's Look at His Word Accuracy – Automatic Evaluation of Tracheoesophageal Speech". In: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 61–64, IEEE Computer Society Press, Philadelphia, USA, 2005.

[Schus 06a] M. Schuster, T. Haderlein, E. Nöth, J. Lohscheller, U. Eysholdt, and F. Rosanowski. "Intelligibility of laryngectomees' substitute speech: automatic speech recognition and subjective rating". *Eur Arch Otorhinolaryngol*, Vol. 263, No. 2, pp. 188–193, 2006.

[Schus 06b] M. Schuster, A. Maier, T. Haderlein, E. Nkenke, U. Wohlleben, F. Rosanowski, U. Eysholdt, and E. Nöth. "Evaluation of Speech Intelligibility for Children with Cleft Lip and Palate by Automatic Speech Recognition". *Int J Pediatr Otorhinolaryngol*, Vol. 70, pp. 1741–1747, 2006.

[Schus 06c] M. Schuster, A. Maier, B. Vogt, E. Nöth, E. Nkenke, A. Marchis, U. Eysholdt, and F. Rosanowski. "Objektive und automatische Ermittlung der Verständlichkeit von Kindern und Jugendlichen mit Lippen-Kiefer-Gaumenspalten". In: M. Gross and F. Kruse, Eds., *Aktuelle phoniatrisch-pädaudiologische Aspekte 2006*, pp. 43–46, Books On Demand GmbH, Norderstedt, Germany, 2006.

[Schus 08] M. Schuster, A. Maier, A. Schützenberger, E. Nkenke, A. Holst, F. Rosanowski, E. Nöth, and U. Eysholdt. "Verständlichkeit von Kindern mit unterschiedlichen orofazialen Spaltfehlbildungen". In: M. Gross and F. Kruse, Eds., *Aktuelle phoniatrisch-pädaudiologische Aspekte 2008*, pp. 158–159, Rheinware Verlag, Mönchengladbach, Germany, 2008.

[Schut 02] H. Schutte and G. Nieboer. "Aerodynamics of esophageal voice production with and without a Groningen voice prosthesis". *Folia Phoniatr Logop*, Vol. 54, pp. 8–18, 2002.

[Sell 01] D. Sell, P. Grunwell, S. Mildinhall, T. Murphy, T. Cornish, D. Bearn, W. Shaw, J. Murray, A. Williams, and J. Sandy. "Cleft Lip and Palate Care in the United Kingdom—The Clinical Standards Advisory Group (CSAG) Study. Part 3: Speech Outcomes". *Cleft Palate-Craniofacial Journal*, Vol. 32, No. 1, pp. 30–37, 2001.

[Shoza 04] M. Shozakai and G. Nagino. "Analysis of Speaking Styles by Two-Dimensional Visualization of Aggregate of Acoustic Models". In: *Proceedings of the International Conference on Speech Communication and Technology (Interspeech)*, pp. 717–720, ISCA, Jeju Island, Korea, 2004.

[Smola 98] A. Smola and B. Schölkopf. "A Tutorial on Support Vector Regression". Tech. Rep., Royal Holloway University of London, 1998. NC2-TR-1998-030.

[Spear 04] C. Spearman. "The Proof and Measurement of Association between Two Things". *Am J Psychol*, Vol. 15, No. 1, pp. 72–101, 1904.

[Stang 71] K. Stange. *Angewandte Statistik II.* Springer Verlag, Berlin, Heidelberg, Germany, 1971.

[Stani 04] P. Stanier and G. Moore. "Genetics of cleft lip and palate: syndromic genes contribute to the incidence of non-syndromic clefts". *Human Molecular Genetics*, Vol. 13, pp. 73–81, 2004.

[Stell 94] A. Stellzig, W. Heppt, and G. Komposch. "Das Nasometer: Ein Instrument zur Objektivierung der Hyperrhinophonie bei LKG-Patienten". *Journal of Orofacial Orthopedics*, Vol. 55, No. 4, pp. 176–180, 1994.

[Stemm 01] G. Stemmer, C. Hacker, E. Nöth, and H. Niemann. "Multiple Time Resolutions for Derivatives of Mel-Frequency Cepstral Coefficients". In: *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE Computer Society Press, Trento, Italy, 2001.

[Stemm 05] G. Stemmer. *Modeling Variability in Speech Recognition.* Logos Verlag, Berlin, Germany, 2005.

[Steve 74] S. Stevens. "Perceptual magnitude and its measurement". In: E. Carterette and M. Friedmann, Eds., *Handbook of Perception (Vol II)*, pp. 361–389, Academic Press, New York, USA, 1974.

[Steve 75] S. Stevens. *Psychophysics.* Wiley, New York, USA, 1975.

[Sturm 08] M. Stürmer, A. Maier, J. Penne, S. Soutschek, C. Schaller, R. Handschu, M. Scibor, and E. Nöth. “3-D Tele-Medical Speech Therapy using Time-of-Flight Technology”. In: *Proceedings of the 4th European Congress For Medical and Biomedical Engineering*, Antwerp, Belgium, 2008. no pagination.

[Tachi 00] T. Tachimura, C. Mori, S. Hirata, and T. Wada. “Nasalance Score Variation in Normal Adult Japanese Speakers of Mid-West Japanese Dialect”. *Cleft Palate-Craniofacial Journal*, Vol. 37, No. 5, pp. 463–467, 2000.

[Teage 90] H. Teager and S. Teager. “Evidence for Nonlinear Production Mechanisms in the Vocal Tract”. In: *Speech Production and Speech Modelling*, pp. 241–261, 1990.

[Tessi 76] P. Tessier. “Anatomical classification of facial, cranifacial, and latero-facial clefts”. *J Maxillofac Surg*, Vol. 4, No. 69, 1976.

[Timmo 01] M. Timmons, R. Wyatt, and T. Murphy. “Speech after repair of isolated cleft palate and cleft lip and palate”. *British Journal of Plastic Surgery*, Vol. 54, pp. 377–384, 2001.

[Tolar 98] M. Tolarova and J. Cervenka. “Classification and birth prevalence of orofacial clefts”. *Am J Med Genet*, Vol. 75, No. 2, pp. 126–137, 1998.

[Vapni 63] V. Vapnik and A. Lerner. “Pattern recognition using generalized portrait method”. *Automation and Remote Control*, Vol. 24, pp. 774–780, 1963.

[Vogt 07] B. Vogt, A. Maier, A. Batliner, E. Nöth, E. Nkenke, U. Eysholdt, and M. Schuster. “Numerische Quantifizierung der Verständlichkeit von Schulkindern mit isolierter und kombinierter Gaumenspalte”. *HNO*, Vol. 55, No. 11, pp. 891–898, 2007.

[Wahls 00] W. Wahlster, Ed. *Verbmobil: Foundations of Speech-to-Speech Translation.* Springer, New York, Berlin, 2000.

[Wanti 02] N. Wantia and G. Rettinger. “The current understanding of cleft lip malformations”. *Facial Plast Surg*, Vol. 18, No. 3, pp. 147–153, 2002.

[Warre 64] D. Warren and A. Dubois. “A Pressure-Flow Technique for Measuring Velopharyngeal Orifice Area During Continuous Speech”. *Cleft Palate J*, Vol. 1, pp. 52–71, 1964.

[Wells 97] J. Wells. *SAMPA computer readable phonetic alphabet*, Chap. Part IV, section B. Mouton de Gruyter, Berlin, Germany, 1997.

[Wendl 05] E. Wendler, W. Seidner, and U. Eysholdt. *Lehrbuch der Phoniatrie und Pädaudiologie.* Georg Thieme Verlag, Stuttgart, New York, 4 Ed., 2005.

[Wendl 86] J. Wendler, A. Rauhut, and H.Krüger. “Classification of Voice Qualities”. *Journal of Phonetics*, Vol. 14, pp. 483–488, 1986.

[Wertz 04] R. T. Wertz, N. Dronkers, and J. Ogar. “Aphasia: The Classical Syndromes”. In: R. D. Kent, Ed., *The MIT Encyclopedia of Communication Disorders*, pp. 249–252, The MIT Press, Cambridge, Massachusetts, USA, 2004.

[White 02a] T. Whitehill. "Assessing Intelligibility in Speakers With Cleft Palate: A Critical Review of the Literature". *Cleft Palate–Craniofacial Journal*, Vol. 39, pp. 50–58, 2002.

[White 02b] T. Whitehill, A. Lee, and J. Chun. "Direct Magnitude Estimation and Interval Scaling of Hypernasality". *Journal of Speech, Language, and Hearing Research*, Vol. 45, pp. 80–88, 2002.

[Wilki 71] J. H. Wilkinson and C. Reinsch. *Linear Algebra, Vol. II of Handbook for Automatic Computation.* Springer-Verlag, New York, USA, 1971.

[Willa 06] E. Willadsen and H. Albrechtsen. "Phonetic Description of Babbling in Danish Toddlers Born With and Without Unilateral Cleft Lip and Palate". *Cleft Palate-Craniofacial Journal*, Vol. 43, No. 2, pp. 189–200, 2006.

[Wilpo 96] J. Wilpon and C. Jacobsen. "A Study of Speech Recognition for Children and the Elderly". In: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 349–352, IEEE Computer Society Press, Atlanta, Georgia, USA, 1996.

[Windr 08] M. Windrich, A. Maier, R. Kohler, E.Nöth, E. Nkenke, U. Eysholdt, and M. Schuster. "Automatic Quantification of Speech Intelligibility of Adults with Oral Squamous Cell Carcinoma". *Folia Phoniatr Logop*, Vol. 60, pp. 151–156, 2008.

[Winds 04] J. Windsor. "Language Disorders in School Age Children: Overview". In: R. D. Kent, Ed., *The MIT Encyclopedia of Communication Disorders*, pp. 326–329, The MIT Press, Cambridge, Massachusetts, USA, 2004.

[Witte 05] I. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann, San Fransisco, CA, USA, 2nd Ed., 2005.

[Wohll 04] U. Wohlleben. *Die Verständlichkeitsentwicklung von Kindern mit Lippen-Kiefer-Gaumen-Segel-Spalten: Eine Längsschnittstudie über spalttypische Charakteristika und deren Veränderung.* Schulz-Kirchner-Verlag, Idstein, Germany, 2004.

[Wurzb 06] T. Wurzbacher, R. Schwarz, M. Döllinger, U. Hoppe, U. Eysholdt, and J. Lohscheller. "Model-based classification of nonstationary vocal fold vibrations". *J Acoust Soc Am*, Vol. 120, No. 2, pp. 1012–1027, 2006.

[Wuyts 00] F. Wuyts, M. D. Bodt, G. Molenberghs, M. Remacle, L. Heylen, B. Millet, K. V. Lierde, J. Raes, and P. V. Heyning. "The Dysphonia Severity Index: An Objective Measure of Vocal Quality Based on a Multiparameter Approach". *Journal of Speech, Language, and Hearing Research*, Vol. 43, pp. 796–809, 2000.

[Wuyts 96] F. Wuyts, M. D. Bodt, L. Bruckers, and G. Molenberghs. "Research Work of the Belgian Study Group on Voice Disorders 1996: Results". *Acta Oto Rhino-Laryngologica Belgica*, Vol. 50, pp. 331–341, 1996.

[Yang 04] J.-H. Yang and Y.-F. Liao. "Unseen Handset Mismatch Compensation based on A Priori Knowledge Interpolation for Robust Speaker Recognition". In: *Proceedings of the International Conference on Speech Communication and Technology (Interspeech)*, pp. 1769–1772, ISCA, Jeju Island, Korea, 2004.

[Yang 99] Y. Yang and X. Liu. “A re-examination of text categorization methods”. In: *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 42–49, Berkley, August 1999.

[Young 01] J. Young, M. O’Riordan, J. Goldstein, and N. Robin. “What Information do Parents of Newborns with Cleft Lip, Palate, or Both Want to Know?”. *Cleft Palate-Craniofacial Journal*, Vol. 38, pp. 55–58, 2001.

[Ysunz 03] A. Ysunza, M. Pamplona, E. Ramírez, S. Canún, M. Sierra, and A. Silva-Rojas. “Videonasopharyngoscopy in Patients with 22q11.2 Deletion Syndrome (Shprintzen syndrome)”. *Int J Pediatr Otorhinolaryngol*, Vol. 67, pp. 911–915, 2003.

[Ysunz 04] A. Ysunza, M. Pamplona, F. Molina, M. Drucker, J. Felemovicius, E. Ramirez, and C. Patiño. “Surgery for speech in cleft palate patients”. *Int J Pediatr Otorhinolaryngol*, Vol. 68, pp. 1499–1505, 2004.

[Ysunz 97] A. Ysunza, C. Pamplona, T. Femat, I. Mayer, and M. Garcia-Velasco. “Videonasopharngoscopy as an instrument for visual biofeedback during speech in cleft palate patients”. *Int J Pediatr Otorhinolaryngol*, Vol. 41, pp. 291–298, 1997.

[Zenne 86] H. Zenner. “The Post-Laryngectomy Telephone Intelligibility Test (PLTT)”. In: I. Herrmann, Ed., *Speech Restoration via Voice Prosthesis*, pp. 148–152, Springer, Berlin, Heidelberg, Germany, 1986.

[Zevce 02] A. Zečević. *Ein sprachgestütztes Trainingssystem zur Evaluierung der Nasalität.* PhD thesis, University Mannheim, Germany, 2002.

[Zimme 80] H. Zimmerman. “OSI Reference Model—The ISO Model of Architecture for Open Systems Interconnection”. *IEEE Transactions on Communications*, Vol. 28, No. 4, pp. 425–432, 1980.

[Zwick 67] E. Zwicker and R. Feldtkeller. *Das Ohr als Nachrichtenempfänger.* Hirzel, Stuttgart, Germany, 1967.

Index