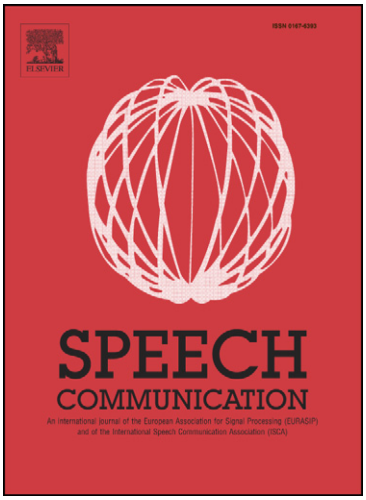
Peaks – A System for the automatic evaluation of voice and speech disorders

A. Maier, T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster, E. Nöth

Accepted Manuscript


ISSN 0167-6393
ELSEVIER
SPEECH
COMMUNICATION
An international journal of the European Association for Signal Processing (EURASIP) and of the International Speech Communication Association (ISCA)

Peaks – A System for the automatic evaluation of voice and speech disorders

A. Maier, T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster, E. Nöth

PII: S0167-6393(09)00003-X
DOI: 10.1016/j.specom.2009.01.004
Reference: SPECOM 1772

To appear in: *Speech Communication*

Received Date: 28 July 2008
Revised Date: 9 December 2008
Accepted Date: 13 January 2009

Please cite this article as: Maier, A., Haderlein, T., Eysholdt, U., Rosanowski, F., Batliner, A., Schuster, M., Nöth, E., Peaks – A System for the automatic evaluation of voice and speech disorders, *Speech Communication* (2009), doi: 10.1016/j.specom.2009.01.004

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

PEAKS – A SYSTEM FOR THE AUTOMATIC EVALUATION OF VOICE AND SPEECH DISORDERS

A. Maier [a,b] T. Haderlein [a,b] U. Eysholdt [a] F. Rosanowski [a]
A. Batliner [b] M. Schuster [a] E. Nöth [b]

[a] *Abteilung für Phoniatrie und Pädaudiologie, Universität Erlangen-Nürnberg*
Bohlenplatz 21, 91054 Erlangen, Germany

[b] *Lehrstuhl für Mustererkennung, Universität Erlangen-Nürnberg*
Martensstraße 3, 91058 Erlangen, Germany

Andreas.Maier@informatik.uni-erlangen.de

Abstract

We present a novel system for the automatic evaluation of speech and voice disorders. The system can be accessed via the internet platform-independently. The patient reads a text or names pictures. His or her speech is then analyzed by automatic speech recognition and prosodic analysis. For patients who had their larynx removed due to cancer and for children with cleft lip and palate we show that we can achieve significant correlations between the automatic analysis and the judgment of human experts in a leave-one-out experiment (p<0.001). A correlation of .90 for the evaluation of the laryngectomees and .87 for the evaluation of the children's data was obtained. This is comparable to human inter-rater correlations.

Key words: Speech intelligibility, speech and voice disorders, automatic evaluation of speech and voice pathologies

1 Introduction

Communication is important for our daily life. About 87.5 % of the inhabitants of urban areas require communication for their daily work. Communication disorders have a major impact 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. 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 (cf. Ruben (2000)). The use of automatic speech processing techniques will contribute to reduce the cost of the care of communication disorders as well provide better rehabilitation of such disorders and hence increase the employment opportunities for people with such disorders.

The first step to achieve this goal is to objectify the assessment of communication disorders in order to reduce manual efforts. For the evaluation of the quality of communication disorders objective methods are necessary at least for the following aspects:

(1) patient assessment:
How severe is the disorder?
(2) therapy control:
Quantification of changes in the disorder
(3) scientific evaluation: Which disease or therapy method shows better overall results when tested on groups of patients?
(4) specification of the disorder and its impact on the communication skills: Automatic quantification of a disorder allows the computation of the relation between a certain disorder and the reduction of the quality of the global speech outcome.

The assessment of communication disorders or intelligibility is usually performed subjectively. Although speech pathologists receive intensive training to ensure the reliability of their ratings, inter-rater correlations often fall below 0.9 on difficult tasks. This is caused by individually differing experience and variable test conditions (Paal et al. (2005); Keuning et al. (1999)). For scientific purposes, evaluation is usually performed by a panel of listeners. Semi-standardized instruments for the analysis of speech disorders in children and adults are well known (Panchal et al. (1996); Paulowski et al. (1998); Mády et al. (2003); Enderby (2004)). These subjective methods are still the most commonly used to assess speech intelligibility (Robbins et al. (1987); Bodin et al. (1994); Brown et al. (1997); Knuuttila et al. (1999); Haughey et al. (2002); Seikaly et al. (2003); Markkanen-Leppanen et al. (2006)), speech disorders and temporal structure of speech (Mahanna et al. (1998); Pauloski et al. (1998); Furia et al. (2001); Su et al. (2003); Bressmann et al. (2004); Terai and Shimahara (2004)). Until now, automatic diagnostic tools for the assessment of speech after treatment have only been performed for single aspects such as the quantification of nasalance as in Kuttner et al. (2003) and spectral characteristics and intensity of the voice signal as in Zečević (2002). However, these methods have limitations and do not allow assessing speech intelligibility in a comprehensive and reliable way.

In this article, we introduce PEAKS (**P**rogram for **E**valuation and **A**nalysis of all **K**inds of **S**peech disorders), a recording and analysis environment for the automatic or manual evaluation of voice and speech disorders. The system can be accessed via the internet or a public telephone, i.e. the system does not require any special hardware except for a standard PC with internet access and a sound card. The patient performs a standardized test which is then automatically rated. Therefore, the system can be employed in specialized centers for voice and speech disorders, e.g. in cleft centers, where a patient's speech is not always judged by the same person. In addition, the system can also provide a speech therapist who works on her own with a second opinion obtained by our automatic evaluation technique at virtually no additional cost. Hence, the system is able to support the therapist with a calibrated opinion which she can take into consideration during her diagnosis. The analysis might help her to identify additional problems she did not notice right from the start. Furthermore, the system can provide reliable information when the patient has to change therapists.

Although the pathologies which are presented here are very different and complex, speech intelligibility is a superordinate parameter of all voice and speech disorders. For the analysis we use an automatic speech recognition (ASR) system and an automatic prosody module. The output of the ASR system is a recognition rate. The outputs of the prosody module are acoustic prosodic features, such as the slope of the fundamental frequency over time and the energy. The result of the analysis is presented to the user and can be compared to previous recordings of the same patient or to recordings from other patients.

The program is evaluated on voice and speech disorders with a wide range of intelligibility on patients who underwent total laryngectomy, due to laryngeal cancer, and on children with cleft lip and palate. Results were compared to the state-of-the-art evaluation — perceptual rating by a panel of experts.

2 Patients and Methods

2.1 Voice Disorder: Tracheoesophageal Substitute Voice

The tracheoesophageal (TE) substitute voice is currently state-of-the-art treatment to restore the ability to speak after laryngectomy, i.e. the total removal of the larynx (after cancer of the larynx, cf. Brown et al. (2003)): A silicone one-way valve is placed into a shunt between the trachea and the esophagus, which on the one hand prevents aspiration and on the other hand deviates the air stream into the upper esophagus while the patient exhales. Tissue vibrations of the pharyngo-esophageal segment modulate the streaming air comparable to laryngeal voice production and generate the primary substitute voice signal which is then further modulated in the same way as normal voice. In comparison to normal voices, the quality of substitute voices is "low". Intercycle frequency perturbations result in a hoarse voice (cf. Schutte and Nieboer (2002)). Furthermore, dynamic parameters such as pitch and volume are restricted which leads to monotone speech. Acoustic studies of TE voices can be found for instance in Robbins et al. (1984) and Bellandese et al. (2001).

2.2 Speech Disorders of Children with Cleft Lip and Palate

Cleft lip and palate (CLP) is the most common malformation of the head with incomplete closure of the cranial vocal tract (cf. Wantia and Rettinger (2002); Millard and Richman (2001); Rosanowski and Eysholdt (2002); Schönweiler and Schönweiler (1994); Schönweiler et al. (1999)). Speech disorders can still be present after reconstructive surgical treatment. The characteristics of speech 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 the back of the tongue or vice versa, and a modified articulatory tension, e.g. weakening of the plosives (cf. Harding and Grunwell (1998)). They affect not only the intelligibility but therewith the social competence and emotional development of a child.

2.3 Speech Material

41 laryngectomees ($\mu = 62.0 \pm 7.7$ years old, 2 female and 39 male) with TE substitute voice read the German version of the text "The North Wind and the Sun", a fable from Aesop. It is a phonetically rich text with 108 words, of which 71 are unique. The speech samples were recorded with a close-talking microphone at 16 kHz sampling frequency and 16 bit resolution.

PEAKS was also applied during the regular out-patient examination of 31 children with CLP (mean 10.1 ± 3.8 years). All children were native German speakers, some of them using a local dialect. Their therapies were performed according to their cleft type and their individual needs. The speech data were recorded using a German standard speech test, the "Psycho-Linguistische Analyse Kindlicher Sprech-Störungen" (Psycho-Linguistic Analysis of Children's Speech Disorders – PLAKSS Fox (2002)). The test consists of 33 slides which show pictograms of the words to be named. In total the test contains 99 words which include all German phonemes in different positions. Additional words, however, were uttered in between the target words since some children tended to explain the pictograms with multiple words. For the transliteration the children's speech was segmented into turns semi-automatically. The data were, therefore, automatically segmented at pauses which were longer than one second. Turns containing the speech of the speech therapist only were removed manually. The therapist's speech was also manually removed from turns containing children's and the therapist's speech. Each of them contains 2.3 words on average. In total 2209 of these utterances were obtained.

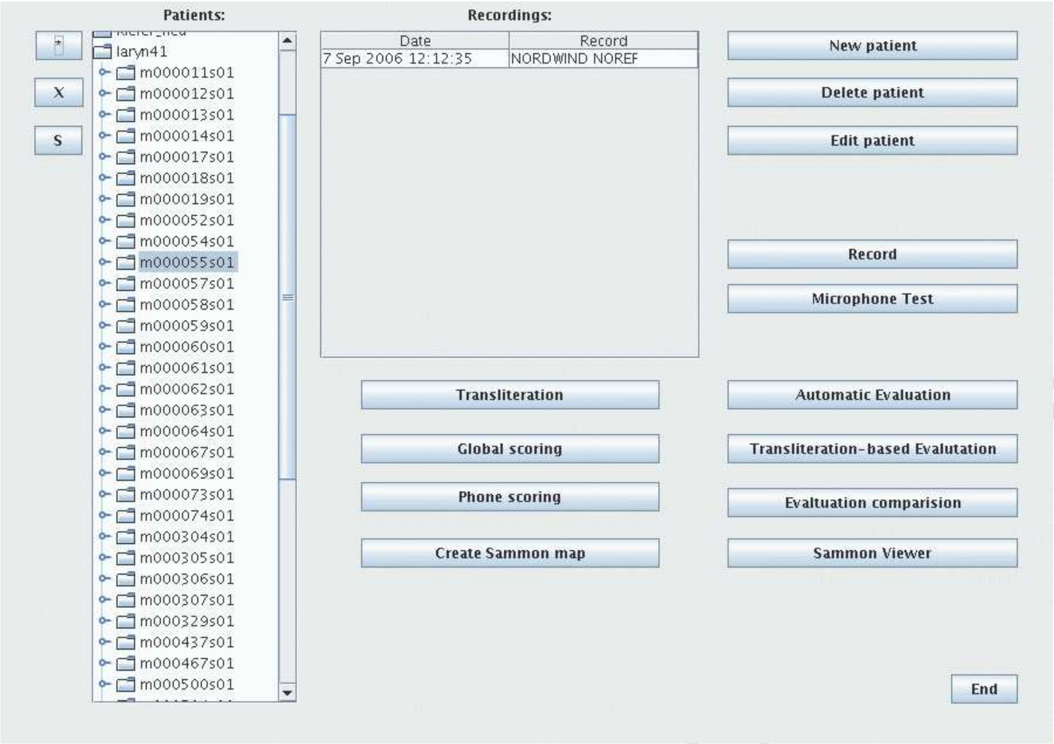
2.4 Perceptive Evaluation

A group of five voice professionals subjectively evaluated both databases while listening to a recording of the speech data. A five-point Likert scale (1 = very high, 2 = rather high, 3 = medium, 4 = rather low, 5 = very low) was applied to rate the intelligibility of each recording. The experts rated the intelligibility in each turn on the same Likert scale as before. In the case of the laryngectomees, the speech was read more or less fluently and therefore their data were listened to in a single turn. In the case of the CLP children, the test data consisted of pictograms which were to be named. In between many of the pictograms long pauses occurred because the children had to think of the correct name for the pictogram. Therefore, the long pauses were automatically detected and removed to speed up the evaluation procedure. This procedure results in an average of 70 turns per child. First, the score for all turns of a patient was averaged for each expert to represent the intelligibility. In a second step the score of each patient was then computed as the average of all five expert scores. In this manner an averaged mark – expressed as a floating point value – for each patient could be calculated.

To judge the agreement between the different raters, we calculated Pearson's and Spearman's correlation coefficients. For each rater we calculated the correlation between her/his intelligibility rating and the average of the 4 other raters.

2.5 The PEAKS Recording environment

For routine use of an evaluation system, it must be easily available from any examination room and inexpensive. We created PEAKS, a client/server recording environment. The system can be accessed from any PC with internet access, a webbrowser, a sound card, and an up-to-date Java Runtime Environment (JRE). A registered physician can group his/her patients according to disorder, create new patient entries, create new recordings, analyze patients and groups of patients (cf. Fig. 1). The physician has only access to the data of "his" patients.



Date	Record
7 Sep 2006 12:12:35	NORDWIND NOREF

Fig. 1. Screenshot of the main menu: On the left side, a list the of patients of the currently logged in physician is shown. If a patient is selected, a list of all of his/her recordings is displayed in the center of the screen. On the right side and the lower middle, buttons for different actions are available.

The texts to be read and pictograms to be named by the patient are displayed in the browser. In Fig. 2 the screen during the recording of the PLAKSS test as described in Fox (2002)[1]. is shown. The patient's utterances are recorded by the client. The recording starts when the pictogram or text passage is displayed, and ends when a button is pressed for the next pictogram or text passage. The speech recording is transferred to the server and then the system analyzes the data. The evaluation results can then be reviewed with the client software. The recordings are stored in an SQL database. A secure connection is used for all data transfer. Instead of the patients' names pseudonyms are used in the system to keep personal data as safe as possible. PEAKS is already being used by different departments of our university clinic for scientific purposes. More information can be found at `http://peaks.informatik.uni-erlangen.de/`.

[1] The PLAKSS test is used with permission from Hartcourt Test Services for scientific purposes only.

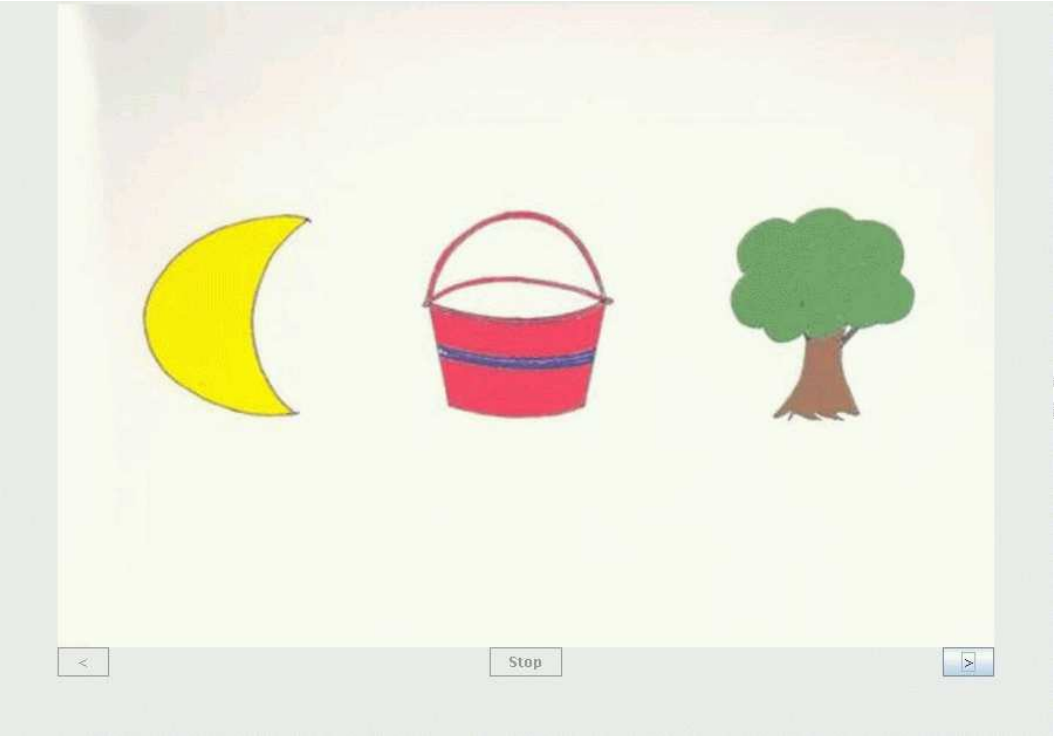


Fig. 2. Screenshot of the recording environment: The picture shows the first slide of the PLAKSS test as described in Fox (2002).

2.5.1 The Automatic Speech Analysis System

For the objective measurement of the intelligibility of pathologic speech, we use an automatic speech recognition system based on Hidden Markov Models (HMM). It is a word recognition system developed at the Chair of Pattern Recognition (Lehrstuhl für Mustererkennung) of the University of Erlangen-Nuremberg. In this study, the latest version as described in detail in Gallwitz (2002) and Stemmer (2005) was used. A commercial version of this recognizer is used in high-end telephone-based conversational dialogue systems by *Sympalog* (www.sympalog.com).

As features we use 11 Mel-Frequency Cepstrum Coefficients (MFCCs) and the energy of the signal plus their first-order derivatives. The short-time analysis applies a Hamming window with a length of 16 ms, the frame rate is 10 ms. The filter bank for the Mel-spectrum consists of 25 triangular filters. The 12 delta coefficients are computed over a context of 2 time frames to the left and the right side (56 ms in total).

The recognition is performed with semi-continuous Hidden Markov Models. The codebook contains 500 full covariance Gaussian densities which are shared by all HMM states. The elementary recognition units are polyphones (Schukat-Talamazzini et al. (1993)), a generalization of triphones. Polyphones use phones in a context as large as possible which can still statistically be modeled well, i.e., the context appears more often than 50 times in the training data. The HMMs for the polyphones have three to four states.

We used a unigram language model to weigh the outcome of each word model. It was trained with the transliteration of the spoken tests, i.e., the vocabulary size was 71 words for the TE data and 99 words for the CLP data. Furthermore, additional 266 additional filler-words and common word alternatives were added to the vocabulary of the CLP data. Thus, the frequency of occurrence for each word in the used text was known to the recognizer. This helps to enhance recognition results by including linguistic information. However, for our purpose it was necessary to put more weight on the recognition of acoustic features. In Riedhammer et al. (2007) a comparison between unigram and zerogram language models was conducted. It was shown that intelligibility can be predicted using word recognition accuracies computed using either zero- or unigram language models. The unigram, however, is computationally more efficient because it can be used to reduce the search space. The use of higher n-gram models was not beneficial.

The result of the recognition is a recognized word chain. In order to get an estimate of the quality of the recognition, two criteria are commonly used. The word recognition rate (WR) and the word accuracy (WA) are both computed from the number of correctly recognized words C and the number of words in the reference R. While the WR is just the percentage of correctly recognized words, i.e.,

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

the WA is additionally weighted with the number or wrongly inserted words I:

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

Hence, the WR is defined between 0 % and 100 % while the WA ranges theoretically between minus infinity and 100 %.

2.5.2 Recognizer Training Data

The basic training set for our recognizer for TE speech are dialogues from the Verbmobil project (Wahlster (2000)). The topic of the recordings is appointment scheduling of normal speakers. The data were recorded with a close-talk microphone with 16 kHz sampling frequency and 16 bit resolution. The speakers were from all over Germany and thus covered most dialect regions. However, they were asked to speak standard German. About 80% of the 578 training speakers (304 male, 274 female) were between 20 and 29 years old, less than 10% were over 40. This is important in view of the test data, because the fact that the average age of our test speakers is more than 60 years may influence the recognition results. A subset of the German Verbmobil data (11,714 utterances, 257,810 words, 27 hours of speech) was used for the training set and 48 utterances (1042 words) for the validation set (the training and validation corpora were the same as in Gallwitz (2002) and Stemmer (2005)).

The speech recognition system for children's speech had been trained with acoustic information from 23 male and 30 female children from a local school who were between 10 and 14 years old (6.9 hours of speech). To make the recognizer more robust, we added data from 85 male and 47 female adult speakers from all over Germany (2.3 hours of spontaneous speech from the Verbmobil project, see above). The adults' data were adapted by vocal tract length normalization as proposed in Stemmer et al. (2003). During training an evaluation set was used that only contained children's speech. MLLR adaptation (cf. Gales et al. (1996); Maier et al. (2006)) with the patients' test data led to further improvement of the speech recognition system for the children's speech.

2.5.3 Calculation of Acoustic-Prosodic Features

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 (such as *long vowel*) can be used by the prosody module (cf. Batliner et al. (2000)).

First, the prosody module extracts so called basic features from the speech signal. These are the energy, the fundamental frequency (F_0) after Bagshaw et al. (1993), and the location of voiced and unvoiced segments in the signal. In a second step, the actual prosodic features are computed to model the prosodic properties of the speech signal. For this purpose 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. The end of a word can be provided by any standard word recognizer, and therefore this point can more easily be 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 1). These features model F_0, energy and duration, e.g. the maximal F_0 in the current word. Fig. 3 shows examples of the F_0 features. In addition, 16 global prosodic features for the whole utterance are calculated (cf. Table 2). They cover each of mean and standard deviation for jitter and shimmer, information on voiced and unvoiced sections. The last global feature is the standard deviation of the fundamental frequency F_0. In order to evaluate pathologic speech on test level, 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 (such as "minimum

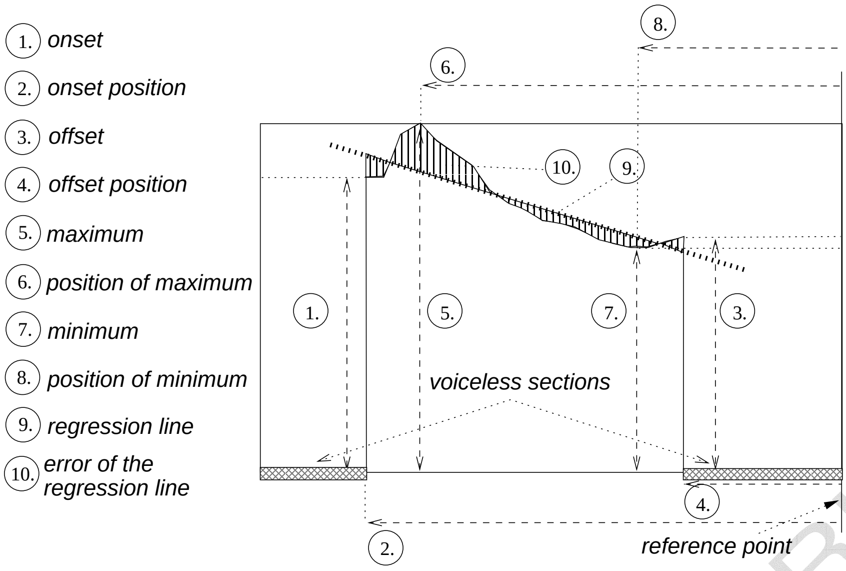


Fig. 3. Computation of prosodic features within one word (after Kießling (1997))

EnergyRegCoeffWord" denoting the minimun energy contour regression slope per word or "maximum F0MeanWord" for the maximum of the mean F_0 per word). Thus we get 148 features for the whole text.

Fig. 3 shows the computation of the word-based F_0 features. The mean values such as F0MeanGlobalWord are computed for a window of 15 words (or less if the utterance is shorter, cf. Batliner et al. (1999, 2001)) so they are regarded as turn-level features here.

In constrast 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', '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 (cf. Batliner et al. (2003a); Huber (2002)), the detection of boundaries between phrases (cf. Batliner et al. (1995)), the user state (cf. Adelhardt et al. (2003); Batliner et al. (2003b)), and the focus of attention (cf. Hacker et al. (2006)).

2.5.4 Feature Selection

In this work we chose the Multiple-Regression/Correlation analysis to determine the best subset with n features. Therefore, we select the feature with the highest correlation to the target criterion as the first feature. Subsequently, we investigate all possible feature sets that can be created by addition of one of the remaining features. Unlike the *Correlation-based Feature Subset* (CFS) selection (cf. Hall (1998)) which computes the unweighted correlation of the features, each subset $\mathcal{S}$ is evaluated using the best weighted combination. Basically the procedure is very similar to the *MAX R* algorithm (Clark, 2004, p.34).

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 than the unweighted process as suggested in Hall (1998).

According to Cohen and Cohen (1983a) the Multiple Regression Analysis can be used for the prediction of y_i using a multidimensional vector $\boldsymbol{x}_i$ with n dimensions:

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

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

$$\boldsymbol{y} = \boldsymbol{c}^\top \boldsymbol{X} \tag{2}$$

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

$$\boldsymbol{c}^\top = \boldsymbol{y}\boldsymbol{X}^* \tag{3}$$

where $\boldsymbol{X}^*$ is the Moore-Penrose pseudo-inverse of $\boldsymbol{X}$ (Moore (1920); Penrose (1955)) 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

$$\hat{y}_i = \boldsymbol{c}^\top \begin{pmatrix} \boldsymbol{x}_i \\ 1 \end{pmatrix}. \tag{4}$$

Since this computation involves a lot of matrix inversions it is quite slow, as can be reviewed in (Clark, 2004, p.34). The matrix inversion has a theoretical complexity of $\mathcal{O}(N_{\mathcal{S}}^3)$ (cf. Press et al. (1992); Courrieu (2005)). Implementations of the matrix inverse based on the QR decomposition like the one in Weka (cf. Witten and Frank (2005)) usually have a complexity of $\mathcal{O}^{\text{R-iter}} = \mathcal{O}(N_{\mathcal{S}}^2 N)$ where N is the number of training vectors and $N_{\mathcal{S}}$ the number of selected features.

We propose faster approximation of R which 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 $\boldsymbol{\Phi}_{\text{FS},\nu}$ — an identity matrix where row ν is removed — with $\boldsymbol{X}$ to remove feature number ν. So the parameters $\boldsymbol{c}_{\mathcal{S}}$ can be computed according to Eqs. 3 and 2:

$$\boldsymbol{c}_{\mathcal{S}}^{\top} = \boldsymbol{y}\boldsymbol{X}_{\mathcal{S}}^{*} = \boldsymbol{y}\boldsymbol{X}^{*}\boldsymbol{\Phi}_{\mathrm{FS},\nu}^{\top} = \boldsymbol{c}^{\top}\boldsymbol{\Phi}_{\mathrm{FS},\nu}^{\top} \tag{5}$$

Note that $\boldsymbol{\Phi}_{\mathrm{FS},\nu}^{\top}$ is the pseudo-inverse of $\boldsymbol{\Phi}_{\mathrm{FS},\nu}$ 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}:

$$\epsilon_{\mathrm{R}}(\boldsymbol{c}_{\mathcal{S}}) = \sum_{i=1}^{N}\left(\boldsymbol{c}^{\top}\boldsymbol{x}_{\boldsymbol{i}} - y_i\right)^2 \tag{6}$$

Differentiation after each component c_j yields the following gradient function:

$$\frac{\delta\epsilon_{\mathrm{R}}}{\delta c_j} = \sum_{i=1}^{N}\left(\boldsymbol{c}^{\top}\boldsymbol{x}_{\boldsymbol{i}} - y_i\right) * 2x_{i,j} \tag{7}$$

Using Eq. 5 as initialization for the gradient descent yields a quite good convergence behavior. In terms of complexity, this procedure surpasses the previous method: Since the sums of Eqs. 6 and 7 require just a single pass in each iteration, the complexity $\mathcal{O}^{\text{R-grad}}$ of this methods is

$$\mathcal{O}^{\text{R-grad}} = \mathcal{O}(N * N_{\mathcal{S}} * 2 * C) \tag{8}$$

where C denotes a constant which corresponds to the number of iterations of the gradient descent. Hence, the feature selection is performed with the gradient descent method in order to speed up the feature selection procedure.

2.5.5 Prediction of Expert Scores

With the previously described features we can now assign scores to the recordings of the patients. To reach this goal we pursue prediction of the expert scores since the class value is a floating point value in our experiments. Therefore, we apply *Support Vector Regression* (cf. Smola and Schölkopf (1998)) since it models outliers very well and robustly. In this manner we predict the numeric human scores of the patients' recordings from the feature vector. For the sake of simplicity we will only describe support vector regression with linear kernel.

The goal of SVR 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:

$$\hat{y}_i = \boldsymbol{w}^{\top}\boldsymbol{x}_i + b \tag{9}$$

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

$$y_i - (\boldsymbol{w}\boldsymbol{x}_{\boldsymbol{i}} + b) \le \epsilon \quad \text{and} \quad (\boldsymbol{w}\boldsymbol{x}_i + b) - y_i \le \epsilon. \tag{10}$$

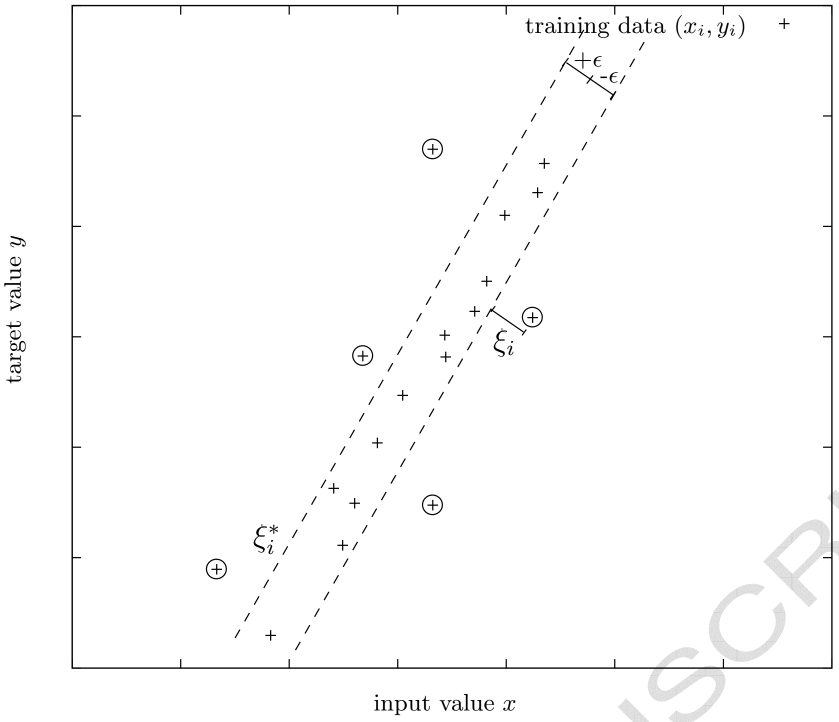


Fig. 4. Support Vector regression finds a function that has at most deviation ϵ from the targets y_i. In order to allow deviations larger than ϵ, a slack variable ξ_i is introduced. Note that the support vectors are outside the ϵ tube.

To allow deviations greater than ϵ, slack variables ξ_i and ξ_i^* are introduced. So Equation 10 can be rewritten to

$$y_i - (\boldsymbol{w}\boldsymbol{x}_i + b) \leq \epsilon + \xi_i \quad \text{and} \quad (\boldsymbol{w}\boldsymbol{x}_i + b) - y_i \leq \epsilon + \xi_i^*. \tag{11}$$

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

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}||w||^2 + C\sum_i(\xi_i + \xi_i^*) \\ \text{subject to} \quad & \begin{cases} y_i - (\boldsymbol{w}\boldsymbol{x}_i + b) \leq \epsilon + \xi_i \\ (\boldsymbol{w}\boldsymbol{x}_i + b) - y_i \leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \end{aligned} \tag{12}$$

Similar as for Support Vector machines as described in Schölkopf (1997), a primal Lagrangian can be formulated introducing Lagrange multipliers α_i, α_i^*, η_i, and η_i^* in order to solve this problem.

$$\begin{aligned} L_P = & \frac{1}{2}||w||^2 + C\sum_i(\xi_i + \xi_i^*) - \sum_i \alpha_i(\epsilon + \xi_i - y_i + \boldsymbol{w}^\top\boldsymbol{x}_i + b) \\ & - \sum_i \alpha_i^*(\epsilon + \xi_i^* + y_i - \boldsymbol{w}^\top\boldsymbol{x}_i - b) - \sum_i(\eta_i\xi_i + \eta_i^*\xi_i^*) \end{aligned} \tag{13}$$

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:

$$0 = \sum_i (\alpha_i^* - \alpha_i) \tag{14}$$

$$\boldsymbol{w} = \sum_i (\alpha_i - \alpha_i^*)\boldsymbol{x}_i \tag{15}$$

$$0 = C - \alpha_i - \alpha_i^* - \eta_i - \eta_i^* \tag{16}$$

By substitution of the equations Eq. 14 to Eq. 16, in Eq. 13 the following optimization problem is obtained:

$$\begin{aligned} &\text{maximize} \begin{cases} -\frac{1}{2}\sum_{i,j}(\alpha_i - \alpha_j)(\alpha_i^* - \alpha_j^*)\boldsymbol{x}_i^\top \boldsymbol{x}_j \\ -\epsilon \sum_i (\alpha_i - \alpha_i^*) + \sum_i y_i(\alpha_i - \alpha_i^*) \end{cases} \\ &\text{subject to} \quad \begin{cases} \sum_i (\alpha_i - \alpha_i^*) = 0 \\ \alpha_i, \alpha_i^* \in [0, C] \end{cases} \end{aligned} \tag{17}$$

Note that the Lagrange multipliers η_i and η_i^* are eliminated in the derivation of Eq. 17. According to Smola and Schölkopf (1998) 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 are both nonzero at the same time. Furthermore, α_i and α_i^* are zero if $|\hat{y}_i - y_i| < \epsilon$. Therefore, *support vectors* can only be found outside the ϵ-tube (cf. Fig. 4). With Eq. 15 the prediction of $\hat{y}_i$ from Eq. 9 can now be written without the actual weight vector $\boldsymbol{w}$:

$$\hat{y}_i = \left[\sum_j (\alpha_j - \alpha_j^*)\boldsymbol{x}_j\right]^\top \boldsymbol{x}_i + b \tag{18}$$

Hence, the predictions can be computed from the *support vectors* $\boldsymbol{x}_j$ without explicit computation of the weight vector $\boldsymbol{w}$.

3 Results

3.1 Perceptual Evaluation

The perceptual evaluation by five experts showed good consistency. Table 3 shows the Pearson (r) and Spearman (ρ) correlation coefficients between each rater and the average correlation coefficients for the TE database. The correlations are all between .77 and .87 and hence in the same range. The 95 % confidence intervals are reported in brackets.

The experts' evaluation showed good consistency on the children's database as well. Table 4 gives an overview on the correlations of the experts to each other. Note that the rating procedure on this database seems to yield higher consistency between the raters in CLP vs. TE speech. We ascribe this effect to the much higher number of ratings (about 70 per child) compared to the low number of ratings (one per laryngectomee) in the other database.

3.2 Automatic Evaluation

Since both corpora contain few speakers, all experiments were performed in a leave-one-out (LOO) manner:

- First, the features are extracted using the speech recognizer and the prosody module.
- Then, the most important features are selected as the subset of n features with the the highest correlation to the target values, i.e. the best linear prediction according to the Multi-Correlation/Regression analysis as described in Cohen and Cohen (1983b) (cf. Section 2.5.4).
- The best feature subset is then used to train a Support Vector Regression which is used to predict the left out value.

These steps are iterated for each speaker. In the end the correlation between the predicted values and the target values is computed in order to determine the prediction accuracy. The number of selected features was increased until the prediction accuracy did not further improve. In this manner prediction-systems are built for the mean of all experts and each single expert.

Because of the fact that the features are selected in every LOO iteration, the new feature sets differ in each iteration. In order to demonstrate which features are of most importance, we report the selected features with the highest mean rank. Note that the feature selection process was performed in a best-first manner. Hence, the mean rank will only change very little if additional features are selected. The occurrence in the list at a high mean rank position does not necessarily guarantee that the feature was selected in every iteration but in many of them.

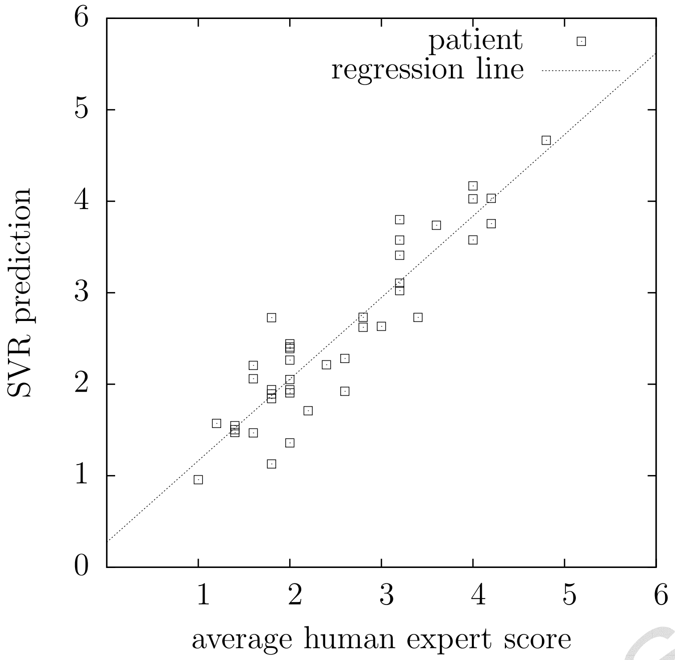


Fig. 5. Predicted expert scores in comparision to the actual expert scores for the laryngectomees case: The LOO predicted value with three features is plotted against the mean expert score.

3.2.1 Automatic Intelligibility Assessment of TE Speech

Table 5 shows correlation coefficients of the LOO prediction on the laryngectomees' database. In all cases the first selected feature to model the intelligibility is either the WA or the WR. Next, prosodic features are added in the feature selection process. We stopped reporting additional features when the correlation did not increase further. Combination of either the WA or the WR with prosodic features yields improvement in most cases. The prediction of the reference—the mean opinion of all experts—is improved by 3.4 % in the case of Pearson's r and 4.8 % for Spearman's ρ relatively. Fig. 5 shows the prediction using three features.

In general the scores of each individual rater are modeled by these features with a correlation of $r > .75$ and $\rho > .73$. The raters 2 and 3 can not be modeled better by further prosodic information. Either the word recognition rate or the word accuracy is already sufficient.

Speech recognition seems to be influenced by the same factors as human perception: The performance of a speech recognition system models the average human perception very well. For two of the raters the recognizer's performance alone was able to model the rater.

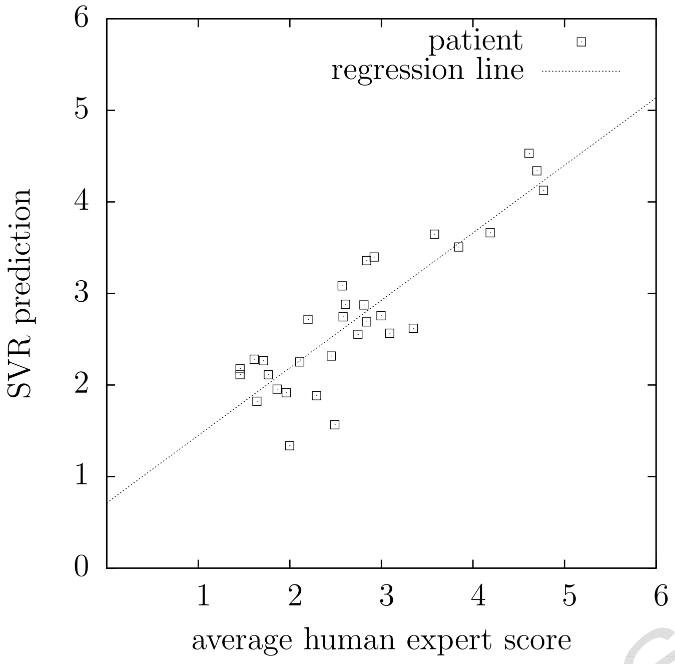


Fig. 6. Predicted expert scores in comparision to the actual expert scores for the CLP children: The LOO predicted value with three features is plotted against the mean expert score.

3.2.2 Automatic Intelligibility Assessment of CLP Speech

The combination of the prosodic features and the result of the speech recognizer is also beneficial for the prediction of experts' scores (cf. Table 6). The best feature for the prediction of the intelligibility is in all raters either the word accuracy or the word recognition rate. The prediction of the mean of all raters is slightly refined 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. Fig. 6 shows the predicted values and the reference in comparision. In terms of Spearman's correlation, the prediction is increased slightly by adding the minimum mean F_0 per word. As shown in Table 6, the selection of the first feature does not yield improvement. The combination with more features, however, helps in the prediction of the experts' scores.

For the prediction of individual experts, only the prediction of one 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 $r > .80$ and a Spearman correlation $\rho > .75$ in this LOO experiment. Again this is in the same range as the experts.

4 Discussion

PEAKS is a system for the evaluation of speech intelligibility. We test the new method on two types of communication disorders: voice and articulation disorders. In both examples the disorder ranges from almost normal intelligibility to severely disordered speech. Both datasets are suitable to show the discriminatory power of our method, because they have a broad range in intelligibility. They are from a completely different origin. However, both inhibit communication, because their effect is a degraded intelligibility.

For comparable results we chose standard evaluation procedures such as a standard text to read or pictograms as usually given for voice and speech evaluation. Another advantage of the procedure is that the vocabulary of the task is limited, which yields a more robust speech recognition.

For the evaluation of the method a valid reference is required. Hence, the data were audited by a panel of speech experts. The procedure yielded a good inter-rater consistency. To reduce the subjectivity in the data, the mean of all experts was computed in order to create a consensus score which we treated to as an appropriate reference (cf. Henningsson et al. (2008)).

In both disorders the PEAKS system agreed with the perceptual evaluation by the speech experts. The correlations were highly significant ($p < 0.001$) and in the same range as the human experts, i.e., within the 95 % confidence interval. In contrast to perceptual evaluation, the evaluation of the system is performed automatically and in less than real time.

Further use of prosodic features was beneficial in the case of the alaryngeal speech. In this manner the quality of the automatic evaluation could be improved. In the case of the children's speech, improvements could only be obtained for individual raters. We relate this to the fact that the children spoke isolated words in most cases. Hence, the impact of the prosody on the intelligibility was only low, i.e., the uttered segments were too short to incorporate a lot of prosody.

In the beginning of this article we postulated four important properties for a system for the automatic assessment of voice and speech disorders. With the previously discussed topics, we now conclude by specifying which of the postulates are covered by our system.

(1) patient assessment: The intelligibility ratings of our evaluation system are in the same range as those of the experts, i.e. our system is suitable for the analysis of voice and speech disorders. Thus, quantification of the disorder can be provided with respect to intelligibility.
(2) therapy control: For a fixed speech input the system produces exactly the

same result when the procedure is repeated once or multiple times, i.e., the intra-rater variability is 0. So the system provides a reliable method for therapy control.

(3) scientific evaluation: With a reliable means of quantification of speech and voice disorders, the system will also be able to evaluate different modes of therapy on the speech outcome in terms of intelligibility.

The fourth point “specification of the disorder and its impact on communication skills”, is yet to be shown by medical studies in the future.

In general, the recording environment is highly suitable for clinical purposes. One major reason is that there are no installation costs, since many examination rooms already provide a PC with internet access. Furthermore, the system is easy to apply. The system is also suitable for screening tests, if age-dependent normative data were acquired.

In the future, we will add more features to the system. First of all we want to add a visualization module to display and compare different patient data sets. The result is a map which displays different disorders in certain regions. Such a visualization could help to classify disorders when the recording of a new patient is projected into a map with well documented patients. Patients who are close to the new patient should be also similar in their disorder.

Furthermore, we want to enable the assessment of distinct speech disorders and voice quality. The assessment result would then be more detailed, i.e., the reduction in intelligibility could then be related to the kind of the disorder. First results are presented in Maier et al. (2008).

5 Conclusion

Our evaluation system provides an easy to apply, cost-effective, instrumental, and objective evaluation of the intelligibility for voice and speech disorders. It is as reliable as human experts.

6 Acknowledgments

This work was funded by the German Cancer Aid (Deutsche Krebshilfe) under grant 106266 and the German Research Foundation (Deutsche Forschungsgemeinschaft) under grant SCHU2320/1-1. The responsibility for the content of this paper lies with the authors. The authors would like to thank both anonymous reviewers of this document for their beneficial suggestions and comments.

References

Adelhardt, J., Shi, R., Frank, C., Zeißler, V., Batliner, A., Nöth, E., Niemann, H., 2003. Multimodal User State Recognition in a Modern Dialogue System. In: the 26th German Conference on Artificial Intelligence. Lecture Notes in Computer Science Springer 2003. Springer, pp. 591–605.

Bagshaw, P., Hiller, S., Jack, M., 1993. 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). ISCA, Berlin, Germany, pp. 1003–1006.
URL `citeseer.ist.psu.edu/169670.html`

Batliner, A., Buckow, A., Niemann, H., Nöth, E., Warnke, V., 2000. The Prosody Module. In: Wahlster, W. (Ed.), Verbmobil: Foundations of Speech-to-Speech Translation. Springer, New York, Berlin, pp. 106–121.

Batliner, A., Buckow, J., Huber, R., Warnke, V., Nöth, E., Niemann, H., 1999. Prosodic Feature Evaluation: Brute Force or Well Designed? In: Proc. of the 14th Intl. Congress of Phonetic Sciences (ICPhS). Vol. 3. San Francisco, USA, pp. 2315–2318.

Batliner, A., Buckow, J., Huber, R., Warnke, V., Nöth, E., Niemann, H., 2001. Boiling down Prosody for the Classification of Boundaries and Accents in German and English. Vol. 4. pp. 2781–2784.

Batliner, A., Fischer, K., Huber, R., Spilker, J., Nöth, E., 2003a. How to Find Trouble in Communication. Speech Communication 40, 117–143.

Batliner, A., Kompe, R., Kießling, A., Nöth, E., Niemann, H., Kilian, U., 1995. The prosodic marking of phrase boundaries: Expectations and Results. In: Rubio, A. (Ed.), New Advances and Trends in Speech Recognition and Coding. NATO ASI Series F. Springer–Verlag, Berlin, pp. 325–328.

Batliner, A., Zeissler, V., Frank, C., Adelhardt, J., Shi, R. P., Nöth, E., 2003b. We are not amused - but how do you know? User states in a multi-modal dialogue system. In: Proceedings of the European Conference on Speech Communication and Technology (Eurospeech). Vol. 1. ISCA, Geneva, Switzerland, pp. 733–736.

Bellandese, M., Lerman, J., Gilbert, H., 2001. An Acoustic Analysis of Excellent Female Esophageal, Tracheoesophageal, and Laryngeal Speakers. J Speech Lang Hear Res 44 (6), 1315–1320.

Bodin, I. K., Lind, M. G., Arnander, C., 1994. Free radial forearm flap reconstruction in surgery of the oral cavity and pharynx: surgical complications, impairment of speech and swallowing. Clin Otolaryngol Allied Sci 19, 28–34.

Bressmann, T., Sader, R., Whitehill, T. L., Samman, N., 2004. Consonant intelligibility and tongue motility in patients with partial glossectomy. J

Oral Maxillofac Surg 62, 298–303.

Brown, D., Hilgers, F., Irish, J., Balm, A., 2003. Postlaryngectomy Voice Rehabilitation: State of the Art at the Millennium. World J Surg 27 (7), 824–831.

Brown, J. S., Zuydam, A. C., Jones, D. C., Rogers, S. N., Vaughan, E. D., 1997. Functional outcome in soft palate reconstruction using a radial forearm free flap in conjunction with a superiorly based pharyngeal flap. Head Neck 19, 524–534.

Clark, V. (Ed.), 2004. SAS/STAT®9.1 User's Guide. SAS Institute, Cary, NC, USA.

Cohen, J., Cohen, P., 1983a. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Cohen, J., Cohen, P., 1983b. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, 2nd Edition. Lawrence Erlbaum Associates, Hillsdale, NJ (USA).

Courrieu, P., 2005. Fast Computation of Moore Penrose Inverse Matrices. Neural Information Processing 8 (2), 25–29.

Enderby, P. M., 2004. Frenchay Dysrarthrie Test. Schulz-Kirchner-Verlag, Idstein, Germany.

Fox, A., 2002. PLAKSS – Psycholinguistische Analyse kindlicher Sprechstörungen. Swets & Zeitlinger, Frankfurt a.M., Germany, now available from Harcourt Test Services GmbH, Germany.

Furia, C. L., Kowalski, L. P., Latorre, M. R., Angelis, E. C., Martins, N. M., Barros, A. P., Ribeiro, K. C., 2001. Speech intelligibility after glossectomy and speech rehabilitation. Arch Otolaryngol Head Neck Surg 127, 877–883.

Gales, M., Pye, D., Woodland, P., 1996. 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). Vol. 3. ISCA, Philadelphia, USA, pp. 1832–1835.

Gallwitz, F., 2002. Integrated Stochastic Models for Spontaneous Speech Recognition. Vol. 6 of Studien zur Mustererkennung. Logos Verlag, Berlin (Germany).

Hacker, C., Batliner, A., Nöth, E., 2006. Are You Looking at Me, are You Talking with Me – Multimodal Classification of the Focus of Attention. In: Sojka, P., Kopeček, I., Pala, K. (Eds.), 9th International Conf. on Text, Speech and Dialogue (TSD). Vol. 4188 of Lecture Notes in Artificial Intelligence. Springer, Berlin, Heidelberg, New York, pp. 581 – 588.

Hall, M. A., 1998. Correlation-based feature subset selection for machine learning. Ph.D. thesis, University of Waikato, Hamilton, New Zealand.

Harding, A., Grunwell, P., 1998. Active versus passive cleft-type speech characteristics. Int J Lang Commun Disord 33 (3), 329–352.

Haughey, B. H., Taylor, S. M., Fuller, D., 2002. Fasciocutaneous flap reconstruction of the tongue and floor of mouth: outcomes and techniques. Arch Otolaryngol Head Neck Surg 128, 1388–1395.

Henningsson, G., Kuehn, D., Sell, D., Sweeny, T., Trost-Cardamone, J., Whitehill, T., 2008. Universal Parameters for Reporting Speech Putcomes in Individuals With Cleft Palate. Cleft Palate Craniofacial Journal 45 (1), 1–17.

Huber, R., 2002. Prosodisch-linguistische Klassifikation von Emotion. Vol. 8 of Studien zur Mustererkennung. Logos Verlag, Berlin, Germany.

Keuning, K., Wieneke, G., Dejonckere, P., 1999. The Intrajudge Reliability of the Perceptual Rating of Cleft Palate Speech Before and After Pharyngeal Flap Surgery: The Effect of Judges and Speech Samples. Cleft Palate Craniofac J 36, 328–333.

Kießling, A., 1997. Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung. Berichte aus der Informatik. Shaker, Aachen, Germany.

Knuuttila, H., Pukander, J., Maatta, T., Pakarinen, L., Vilkman, E., 1999. Speech articulation after subtotal glossectomy and reconstruction with a myocutaneous flap. Acta Otolaryngol 119, 621–626.

Kuttner, C., Schönweiler, R., Seeberger, B., Dempf, R., Lisson, J., Ptok, M., 2003. Normal nasalance for the German language. Nasometric values for clinical use in patients with cleft lip and palate. HNO 51, 151–156.

Mády, K., Sader, R., Hoole, P. H., Zimmermann, A., Horch, H. H., 2003. Speech evaluation and swallowing ability after intra-oral cancer. Clin Linguist Phon 17, 411–420.

Mahanna, G. K., Beukelman, D. R., Marshall, J. A., Gaebler, C. A., Sullivan, M., 1998. Obturator prostheses after cancer surgery: an approach to speech outcome assessment. Prosthet Dent 79, 310–316.

Maier, A., Haderlein, T., Nöth, E., 2006. Environmental Adaptation with a Small Data Set of the Target Domain. In: Sojka, P., Kopeček, I., Pala, K. (Eds.), 9th International Conf. on Text, Speech and Dialogue (TSD). Vol. 4188 of Lecture Notes in Artificial Intelligence. Springer, Berlin, Heidelberg, New York, pp. 431–437.

Maier, A., Hönig, F., Hacker, C., Schuster, M., Nöth, E., 2008. 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.

Markkanen-Leppanen, M., Isatalo, E., Makitie, A. A., Asko-Seljavaara, S., Pessi, T., Suominen, E., Haapanen, M. L., 2006. Changes in articulatory proficiency following microvascular reconstruction in oral or oropharyngeal cancer. Oral Oncol 42, 646–652.

Millard, T., Richman, L., 2001. Different cleft conditions, facial appearance, and speech: relationship to psychological variables. Cleft Palate Craniofac J 38, 68–75.

Moore, E. H., 1920. On the reciprocal of the general algebraic matrix. Bulletin of the American Mathematical Society 26, 394–395.

Paal, S., Reulbach, U., Strobel-Schwarthoff, K., Nkenke, E., Schuster, M., 2005. Beurteilung von Sprechauffälligkeiten bei Kindern mit Lippen-Kiefer-Gaumen-Spaltbildungen. J Orofac Orthop 66 (4), 270–278.

Panchal, J., Potterton, A. J., Scanlon, E., McLean, N. R., 1996. An objective assessment of speech and swallowing following free flap reconstruction for oral cavity cancers. Br J Plast Surg 49, 363–369.

Pauloski, B. R., Rademaker, A. W., Logemann, J. A., Colangelo, L. A., 1998. Speech and swallowing in irradiated and nonirradiated postsurgical oral cancer patients. Otolaryngol Head Neck Surg 118, 616–624.

Paulowski, B. R., Logemann, J. A., Colangelo, L. A., Rademaker, A. W., McConnel, F. M., Heiser, M. A., Cardinale, S., Shedd, D., Stein, D., Beery, Q., Myers, E., Lewin, J., Haxer, M., Esclamado, R., 1998. Surgical variables affecting speech in treated patients with oral and oropharyngeal cancer. Laryngoscope 108, 908–916.

Penrose, R., 1955. A generalized inverse for matrices. Proceedings of the Cambridge Philosophical Society 51, 406–413.

Press, W., Teukolsky, S., Vetterling, W., Flannery, B., 1992. Numerical Recipes in C. Cambridge University Press, Cambridge, MA, USA.

Riedhammer, K., Stemmer, G., Haderlein, T., Schuster, M., Rosanowski, F., Nöth, E., Maier, A., 2007. Towards Robust Automatic Evaluation of Pathologic Telephone Speech. In: Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE Computer Society Press, Kyoto, Japan, pp. 717–722.

Robbins, J., Fisher, H., Blom, E., Singer, M., 1984. A Comparative Acoustic Study of Normal, Esophageal, and Tracheoesophageal Speech Production. J Speech Hear Disord 49 (2), 202–210.

Robbins, K. T., Bowman, J. B., Jacob, R. F., 1987. Postglossectomy deglutitory and articulatory rehabilitation with palatal augmentation prostheses. Arch Otolaryngol Head Neck Surg 113, 1214–1218.

Rosanowski, F., Eysholdt, U., 2002. Phoniatric aspects in cleft lip patients. Facial Plast Surg 18 (3), 197–203.

Ruben, R., 2000. Redefining the survival of the fittest: communication disorders in the 21st century. Laryngoscope 110 (2), 241–245.

Schölkopf, B., 1997. Support vector learning. Ph.D. thesis, Technische Universität Berlin, Germany.

Schönweiler, R., Lisson, J., Schönweiler, B., Eckardt, A., Ptok, M., Trankmann, J., Hausamen, J., 1999. 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 50 (3), 205–217.

Schönweiler, R., Schönweiler, B., 1994. Hörvermögen und Sprachleistungen bei 417 Kindern mit Spaltfehlbildungen. HNO 42 (11), 691–696.

Schukat-Talamazzini, E., Niemann, H., Eckert, W., Kuhn, T., Rieck, S., 1993. Automatic Speech Recognition without Phonemes. In: Proc. European Conf. on Speech Communication and Technology (Eurospeech). Vol. 1. Berlin (Germany), pp. 129–132.

Schutte, H., Nieboer, G., 2002. Aerodynamics of esophageal voice production with and without a Groningen voice prosthesis. Folia Phoniatr Logop 54 (1), 8–18.

Seikaly, H., Rieger, J., Wolfaardt, J., Moysa, G., Harris, J., Jha, N., 2003. Functional outcomes after primary oropharyngeal cancer resection and reconstruction with the radial forearm free flap. Laryngoscope 113, 897–904.

Smola, A., Schölkopf, B., 1998. A tutorial on support vector regression. Tech. rep., Royal Holloway University of London, nC2-TR-1998-030.

Stemmer, G., 2005. Modeling Variability in Speech Recognition. Vol. 19 of Studien zur Mustererkennung. Logos Verlag, Berlin (Germany).

Stemmer, G., Hacker, C., Steidl, S., Nöth, E., 2003. Acoustic Normalization of Children's Speech. In: Proc. European Conf. on Speech Communication and Technology. Vol. 2. Geneva, Switzerland, pp. 1313–1316.

Su, W. F., Hsia, Y. J., Chang, Y. C., Chen, S. G., Sheng, H., 2003. Functional comparison after reconstruction with a radial forearm free flap or a pectoralis major flap for cancer of the tongue. Otolaryngol Head Neck Surg 128, 412–418.

Terai, H., Shimahara, M., 2004. Evaluation of speech intelligibility after a secondary dehiscence operation using an artificial graft in patients with speech disorders after partial glossectomy. Br J Oral Maxillofac Surg 42, 190–194.

Wahlster, W. (Ed.), 2000. Verbmobil: Foundations of Speech-to-Speech Translation. Springer, Berlin (Germany).

Wantia, N., Rettinger, G., 2002. The current understanding of cleft lip malformations. Facial Plast Surg 18 (3), 147–153.

Witten, I., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition. Morgan Kaufmann, San Fransisco, CA, USA.

Zečević, A., 2002. Ein sprachgestütztes Trainingssystem zur Evaluierung der Nasalität. Ph.D. thesis, University Mannheim, Germany.

Table 1
Overview on the prosodic features computed on word level

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 2
Overview on the prosodic features computed on turn level

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 3
Correlation coefficients between single raters and the average of the 4 other raters for the criterion "intelligibility". The 95 % confidence intervals are reported in brackets.

laryngectomees		
rater	mean of other raters	
	r	ρ
rater 1	.84 [.69-.92]	.82 [.66-.91]
rater 2	.87 [.75-.93]	.84 [.69-.92]
rater 3	.80 [.62-.90]	.77 [.57-.88]
rater 4	.81 [.64-.90]	.83 [.68-.91]
rater 5	.80 [.62-.90]	.77 [.57-.88]

Table 4
Correlation coefficients between single raters and the average of the 4 other raters for the criterion "intelligibility". The 95 % confidence intervals are reported in brackets.

children		
rater	mean of other raters	
	r	ρ
rater 1	.94 [.87-.97]	.93 [.84-.97]
rater 2	.94 [.87-.97]	.92 [.82-.96]
rater 4	.94 [.87-.97]	.93 [.84-.97]
rater 6	.95 [.89-.97]	.92 [.82-.96]
rater 7	.96 [.91-.98]	.92 [.82-.96]

Table 5
Correlation between reference ratings and the ratings estimated using SVR using different feature sets on the laryngectomees' database: An increase in the number of features yields an increase in performance in most cases. Due to the LOO procedure only the features with the highest mean ranks are reported. Combination of features is indicated as "+".

feature	prediction SVR		reference raters
	r	ρ	
word accuracy	.87	.85	all raters
+ mean F0MeanWord	**.90**	.87	all raters
+ variance F0OffsetPosWord	**.90**	**.88**	all raters
word recognition rate	.66	.67	rater 1
+ max F0OffsetPosWord	.73	.75	rater 1
+ max PauseSilenceBeforeWord	**.74**	**.76**	rater 1
word recognition rate	**.79**	**.78**	rater 2
word accuracy	**.79**	**.81**	rater 3
word accuracy	**.74**	.77	rater 4
+ variance F0OffsetPosWord	.69	.73	rater 4
+ mean F0MeanWord	.71	.75	rater 4
+ mean PauseSilenceBeforeWord	.69	.73	rater 4
+ mean F0MaxPosWord	**.74**	**.79**	rater 4
word accuracy	.76	.73	rater 5
+ min F0MinWord	**.80**	**.76**	rater 5

Table 6
Correlation between reference rating and the ratings estimated using SVR using different feature sets on the children's database: An increase in the number of features yields an increase in performance only with respect to all raters. Due to the LOO procedure only the features with the highest mean ranks are reported. Combination of features is indicated as "+".

feature	prediction SVR		reference raters
	r	ρ	
word accuracy	.86	.84	all raters
+ minimum EnergyRegCoeffWord	.86	.82	all raters
+ mean Mean_shimmer	**.87**	.82	all raters
+ minimum F0MeanWord	.85	**.87**	all raters
word accuracy	**.83**	**.78**	rater 1
word recognition rate	**.82**	**.79**	rater 2
word accuracy	.82	.80	rater 4
+ minimum F0MaxWord	**.84**	**.86**	rater 4
word accuracy	**.85**	**.83**	rater 6
word accuracy	**.84**	**.81**	rater 7