Intelligibility is more than a single Word: Quantification of Speech Intelligibility by ASR and Prosody

Andreas Maier[1,3], Tino Haderlein[1], Maria Schuster[1], Emeka Nkenke[2], Elmar Nöth[3] ⋆

[1]Abteilung für Phoniatrie und Pädaudiologie, Universität Erlangen-Nürnberg
Bohlenplatz 21, 91054 Erlangen, Germany
[2]Mund-, Kiefer- und Gesichtschirurgische Klinik, Universität Erlangen-Nürnberg,
Glückstraße 11, 91054 Erlangen, Germany
[3]Lehrstuhl für Mustererkennung, Universität Erlangen-Nürnberg
Martensstraße 3, 91058 Erlangen, Germany
Andreas.Maier@informatik.uni-erlangen.de

Abstract. In this paper we examine the quality of the prediction of intelligibility scores of human experts. Furthermore, we investigate the differences between subjective expert raters who evaluated speech disorders of laryngectomees and children with cleft lip and palate. We use the recognition rate of a word recognizer and prosodic features to predict the intelligibility score of each individual expert. For each expert and the mean opinion of all experts we present the best features to model their scoring behavior according to the mean rank obtained during a 10-fold cross-validation. In this manner all individual speech experts were modeled with a correlation coefficient of at least $r > .75$. The mean opinion of all raters is predicted with a correlation of $r = .90$ for the laryngectomees and $r = .86$ for the children.

1 Introduction

Until now speech disorders are evaluated subjectively by an expert listener showing only restricted reliability. For scientific purposes therefore a panel of several expert listeners is needed. For the objective evaluation we developed a new method to quantify speech disorders. In our recent work we evaluated our method with patients whose larynx was removed (laryngectomees) and children with cleft lip and palate (CLP).

By removal of the larynx the patient looses the ability to speak. The patient's breathing is maintained by a detour of the trachea to a hole in the throat—the so called tracheostoma. In order to restore the speech ability of the patient a shunt valve is placed between the trachea and the esophagus. Closure of the tracheostoma forces the air stream from the patient's lungs through the esophagus into the vocal tract. In this way, a tracheoesophageal voice is formed. In

⋆ This work was supported by the Johannes-und-Frieda-Marohn Stiftung and the Deutsche Forschungsgemeinschaft (German Research Foundation) under grant SCHU2320/1-1.

published in: Text, Speech and Dialogue (TSD 2007), Proceedings, LNAI 4629, Springer, Berlin, Heidelberg, 2007, pp. 278-285

comparison to normal voices the quality of such a voice is low [1]. Nevertheless, it is considered as state-of-the-art of substitute voices.

Children with cleft lip and palate suffer from various graduations of speech disorders. The characteristics of these speech disorders are mainly a combination of different articulatory features, e.g. nasal air emissions that lead to nasality, a shift in localization of articulation (e.g. using a /d/ instead of a /g/ or vice versa), and a modified articulatory tension (e.g. weakening of the plosives /t/, /k/, /p/) [2].

In [1] it was shown that—next to the recognition rate of a speech recognizer—prosodic features also hold information on the intelligibility. In this paper we successfully combine both approaches to enhance the prediction quality of our automatic evaluation system for speech disorders. Furthermore, we investigate the individual differences in intelligibility perception and their relation to the prosodic information.

2 Databases

The 41 laryngectomees (mean 62.0 $\pm$ 7.7 years) with tracheoesophageal substitute voice read the German version of the fable "The North Wind and the Sun". It is phonetically balanced and contains 108 words of which 71 are unique.

The children's speech data was recorded using a German standard speech test (PLAKSS [3]). 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 (beginning, center and end of a word). Additional words, however, were uttered in between the target words, since the children tend to explain the pictograms with multiple words. Informed consent had been obtained by all parents of the children prior to the examination. The database contains speech data of 31 children and adolescents with CLP (mean 10.1 $\pm$ 3.8 years).

All speech samples were recorded with a close-talking microphone (DNT Call 4U Comfort headset) at a sampling frequency of 16 kHz and quantized with 16 bit. The data were recorded during the regular out-patient examination of the patients. All patients were native German speakers, some of them using a local dialect.

3 Subjective Evaluation

Both corpora were evaluated by a panel of five speech experts. The experts rated each turn on a Likert scale between 1 $\equiv$ very good and 5 $\equiv$ very bad. So a floating point value was computed for each patient to represent his intelligibility, as commonly used for scientific purposes.

In order to compare the scores we computed Pearson's product moment correlation coefficient r and Spearman's correlation coefficient ρ. Table 1 shows the agreement of the individual raters to mean of the respective other raters.

published in: Text, Speech and Dialogue (TSD 2007), Proceedings, LNAI 4629, Springer, Berlin, Heidelberg, 2007, pp. 278-285

Table 1. Correlations of the individual raters to the mean of the other raters

laryngectomees		
rater	mean of other raters	
	r	ρ
rater L	.84	.82
rater S	.87	.84
rater F	.80	.77
rater K	.81	.83
rater H	.80	.77

children		
rater	mean of other raters	
	r	ρ
rater B	.95	.92
rater K	.94	.93
rater L	.94	.93
rater S	.94	.92
rater W	.96	.92

4 Automatic Speech Recognition System

A word recognition system developed at the `(deleted)` was used. As features we use mel-frequency cepstrum coefficients 1 to 11 plus the energy of the signal for each 16 ms frame (10 ms frame shift). Additionally 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 (HMMs). The codebook contains 500 full covariance Gaussian densities which are shared by all HMM states. The elementary recognition units are polyphones [4]. The polyphones were constructed for each sequence of phones which appeared more than 50 times in the training set.

For our purpose it is necessary to put more weight on the recognition of acoustic features. So we used only a unigram language model to restrict the amount of linguistic information which is used to prune the search tree.

The training set for the adults' speech recognizer are dialogues from the VERBMOBIL project [5]. The topic of the recordings is appointment scheduling. The data were recorded with a close–talking microphone with 16 kHz and 16 bit. The speakers were from all over Germany, and thus covered most regions of dialect. 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 average age of our test speakers is over 60 years; this 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 corpus was the same as in [6]).

The training set of the children's recognizer contained 53 children with normal speech between 10 and 14 years of age. In order to increase the amount of training data, speech data of adult speakers from VERBMOBIL—whose vocal tract length was adapted to children's speech—were added. Further enhancement of the children's recognizer was done by MLLR adaptation to each speaker as described in [7]. A more detailed description of the recognizer, the training set, and the language model is presented in [8, 9].

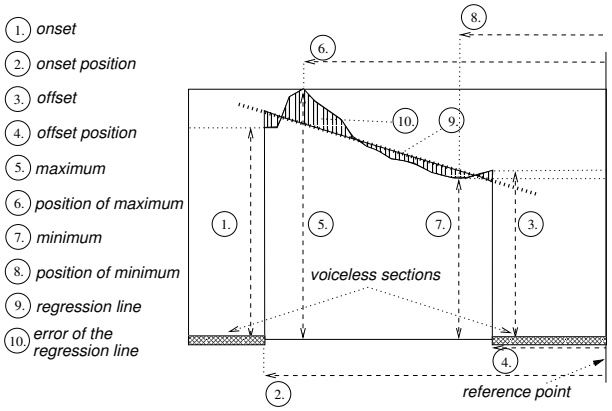

Fig. 1. *Computation of prosodic features within one word (after [10])*

5 Prosodic Features

The prosody module used in these experiments was originally developed within the VERBMOBIL project [5], mainly to speed up the linguistic analysis [11, 12]. It assigns a vector of prosodic features to each word in a word hypothesis graph which is then used to classify a word w.r.t., e.g. carrying the phrasal accent and being the last word in a phrase. For this paper, the prosody module takes the text reference and the audio signal as input and returns 37 prosodic features for each word and then calculates the mean, the maximum, the minimum, and the variance of these features for each speaker, i.e. the prosody of the whole speech of a speaker is characterized by a 148-dimensional vector. These features differ in the manner in which the information is combined (cf. Fig. 1):

1. onset
2. onset position
3. offset
4. offset position
5. maximum
6. position of maximum
7. minimum
8. position of minimum
9. regression line
10. mean square error of the regression line

These features are computed for the fundamental frequency (F_0) and the energy (absolute and normalized). Additional features are obtained from the duration and the length of pauses before and after the respective word. Furthermore jitter, shimmer and the length of voiced (V) and unvoiced (UV) segments are calculated as prosodic features.

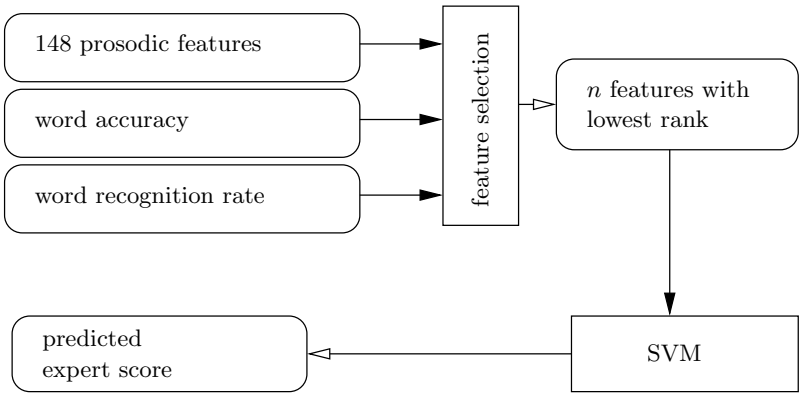

Fig. 2. Proposed system for the prediction of the expert scores

6 Automatic Evaluation

The automatic evaluation system employs support vector regression (SVR) [13] for prediction of the experts' scores.

As displayed in Fig. 2 we utilize on the one hand the word accuracy (WA) and the word recognition rate (WR) of a speech recognizer.

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

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

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

weights the number of wrongly inserted words I in this percentage.

On the other hand 148 prosodic features as features for the system. So we obtain 150 features in total. In order to select a subset of the features we applied a simple algorithm based on the multiple regression/correlation analysis [14] (also called "linear regression" in some cases). 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). This algorithm returned better features than other feature selection algorithms like correlation-based feature subset selection [15] or consistency subset evaluation [16]. 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 10-fold cross validation (CV) 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 CV iterations for each feature. This, however, does not mean, that the particular feature has been selected for all CV iterations.

Table 2. Overview on the prediction performance done by different feature sets on the laryngectomees' database

feature	mean rank	prediction SVR		reference raters
		r	ρ	
word accuracy	0	.87	.83	all raters
mean F_0 of all words	17	**.90**	**.87**	all raters
word recognition rate	17.1	.68	.67	rater L
word accuracy	18.6	.70	.71	rater L
maximum silence before word	34.7	.75	.76	rater L
mean F_0 regression line	39	**.77**	**.78**	rater L
word recognition rate	14.9	**.77**	**.74**	rater S
word accuracy	14.9	**.76**	**.78**	rater R
word accuracy	16.9	.71	.72	rater K
mean silence after word	23.2	.69	.73	rater K
maximum F_0 minimum position	35.5	.74	.78	rater K
maximum F_0 minimum	46.6	**.77**	**.78**	rater K
word accuracy	0	.76	.70	rater H
minimum F_0 minimum	30.6	**.78**	**.72**	rater H

7 Results

The additional use of prosody could enhance the accuracy of the prediction compared to [1] and [9] for the adults' speech data.

Table 2 gives an overview about the quality of the CV prediction on the laryngectomees' database. We stopped reporting additional features when the correlation did not increase further. Combination of either the WR or the WA with prosodic features yields improvement in most cases. The prediction of our gold standard—the mean opinion of all experts—is improved by 3.4 % in case of Pearson's r and 4.8 % for Spearman's ρ relatively. Note that the correlations cannot be compared directly to those of Table 1 since these correlations were not computed in cross-validated manner.

Furthermore, prosody is also useful to model the intelligibility perception of each individual expert. As can be seen in Table 2 rater L's intelligibility scores are modeled best by the

- word recognition rate,
- the word accuracy,
- the maximum silence before each word, and the
- mean of the of the regression coefficient of the fundamental frequency's slope.

The scores of each individual rater are modeled by these features with a correlation of $r > .75$ and $\rho > .72$. The raters S and R seem to judge the intelligibility only by means of either WR or WA. Their opinion of intelligibility cannot be explained further by means of prosody.

With the children's data no further improvement was obtained by the application of prosodic features for the prediction of experts' scores (cf. Table 3).

Table 3. Prediction of the experts' scores by different feature sets on the children's database

feature	mean rank	prediction SVR		reference raters
		r	ρ	
word accuracy	0	**.86**	**.84**	all raters
word accuracy	0	**.86**	**.84**	rater B
word accuracy	14.9	**.77**	**.76**	rater K
word accuracy	0	**.78**	**.77**	rater L
word recognition rate	36.4	**.80**	**.77**	rater S
word accuracy	0	**.80**	**.80**	rater W

Although high correlations between single prosodic features and the mean opinion of the experts exist, the best feature for the prediction of the experts is always the word accuracy.

We suppose that the summarization of the prosodic information is too rough for the case of the children. For the laryngectomees the mean, the variance, the minimum, and the maximum of the prosody of all single words seems to be enough. This might be related to the kind of this disorder: most affected are the fundamental frequency and the duration of pauses since the generation of the tracheoesophageal speech is artifical and the speaking with such a voice is exhausting and the speaker has to stop more often and unexpectedly. Both effects seem to reduce the intelligibility.

For the case of children the prosody of the single words of the speech test has to be differentiated more closely: The prosody of the children depends on the difficulty of the target words. We assume that the prosody of the difficult words is more monotonous than the prosody of familiar and simple words, which are also uttered in between the target words. We will examine this aspect more closely in our future work.

8 Summary

In this paper we successfully combined prosodic features with the recognition rate of a word recognizer to improve the reliability of the automatic speech intelligibility quantification system. A feature selection using multiple regression analysis yielded a prediction system that computes scores which are very close to the experts' scores (r =.90 and ρ =.87). For the data of the children no further improvement was obtained by additional prosodic information. However, the quality of the children's prediction system (r =.87 and ρ =.84) is in the same range as the laryngectomees' prediction system. Therefore, both systems can be used to replace the time and cost intensive subjective evaluations. In addition, the system can be used to investigate the intelligibility perception of the human experts. So a list of features can be computed which models the intelligibility rating of each expert best. The prediction of these single expert models has always a correlation which is above $r > .75$ and $\rho > .74$.

References

1. T. Haderlein, E. Nöth, M. Schuster, U. Eysholdt, and F. Rosanowski, "Evaluation of Tracheoesophageal Substitute Voices Using Prosodic Features," in *Proc. Speech Prosody, 3rd International Conference*, R. Hoffmann and H. Mixdorff, Eds., Dresden, Germany, 2006, pp. 701–704, TUDpress.
2. A. Harding and P. Grunwell, "Active versus passive cleft-type speech characteristics," *Int J Lang Commun Disord*, vol. 33, no. 3, pp. 329–52, 1998.
3. A. Fox, "PLAKSS - Psycholinguistische Analyse kindlicher Sprechstörungen," Swets & Zeitlinger, Frankfurt a.M., now available from Harcourt Test Services GmbH, Germany, 2002.
4. E.G. Schukat-Talamazzini, H. Niemann, W. Eckert, T. Kuhn, and S. Rieck, "Automatic Speech Recognition without Phonemes," in *Proceedings European Conference on Speech Communication and Technology (Eurospeech)*, Berlin, Germany, 1993, pp. 129–132.
5. W. Wahlster, Ed., *Verbmobil: Foundations of Speech-to-Speech Translation*, Springer, New York, Berlin, 2000.
6. G. Stemmer, *Modeling Variability in Speech Recognition*, Ph.D. thesis, Chair for Pattern Recognition, University of Erlangen-Nuremberg, Germany, 2005.
7. M. Gales, D. Pye, and P. Woodland, "Variance compensation within the MLLR framework for robust speech recognition and speaker adaptation," in *Proc. ICSLP '96*, Philadelphia, USA, 1996, vol. 3, pp. 1832–1835.
8. 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*, Hong Kong, China, 2006, vol. 4, pp. 274–277.
9. 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.
10. A. Kießling, *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung*, Berichte aus der Informatik. Shaker, Aachen, 1997.
11. E. Nöth, A. Batliner, A. Kießling, R. Kompe, and H. Niemann, "Verbmobil: The Use of Prosody in the Linguistic Components of a Speech Understanding System," *IEEE Trans. on Speech and Audio Processing*, vol. 8, pp. 519–532, 2000.
12. A. Batliner, A. Buckow, H. Niemann, E. Nöth, and V. Warnke, "The Prosody Module," In Wahlster [5], pp. 106–121.
13. A. Smola and B. Schölkopf, "A tutorial on support vector regression," in *NeuroCOLT2 Technical Report Series*. 1998, NC2-TR-1998-030.
14. J. Cohen and P. Cohen, *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1983.
15. M. A. Hall, *Correlation-based Feature Subset Selection for Machine Learning*, Ph.D. thesis, University of Waikato, Hamilton, New Zealand, 1998.
16. H. Liu and R. Setiono, "A probabilistic approach to feature selection - a filter solution," in *13th International Conference on Machine Learning*, 1996, pp. 319–327.