Automatic Scoring of the Intelligibility in Patients with Cancer of the Oral Cavity

Andreas Maier[1,2]*, Maria Schuster*[1]*, Anton Batliner*[2]*, Elmar Nöth*[2]*, Emeka Nkenke*[3]

[1] Abteilung für Phoniatrie und Pädaudiologie, Universität Erlangen–Nürnberg, Bohlenplatz 21, 91054 Erlangen, Germany
[2] Lehrstuhl für Mustererkennung, Universität Erlangen–Nürnberg, Martensstraße 3, 91058 Erlangen, Germany
[3] Mund- Kiefer– und Gesichtschirurgische Klinik der Universität Erlangen–Nürnberg Glückstr. 11, 91054 Erlangen, Germany

Andreas.Maier@informatik.uni-erlangen.de

Abstract

After surgical treatment of cancer of the oral cavity patients often suffer from functional restrictions such as speech disorders. In this paper we present a novel approach to assess the outcome of the treatment w.r.t. the intelligibility of the patient using the result of an automatic speech recognition system. The word recognition rate was taken as intelligibility score. Compared to four speech experts this method yields results that are as good as the best speech expert compared to the other experts. The correlation between our system and the mean opinion of the experts is .92. Furthermore we show that our system has better performance than the average expert and is more reliable.

Index Terms: Speech intelligibility, Speech processing, Biomedical acoustics, Acoustic applications

1. Introduction

Cancer of the oral cavity is one of the ten most common malignant diseases of humans and is mostly treated by surgery and radiation, sometimes combined with chemotherapy which cause morphologic changes. As a consequence functional disorders such as nutrition and speech disorders occur. Until now, the latter was usually evaluated by perceptive rating performed by an expert. Semi-standardized instruments for the analysis of speech disorders in adults are well known [1, 2, 3, 4, 5]. Yet, the assessment of speech disorders or intelligibility is usually performed subjectively; it therefore lacks reliability because of individually differing experience and variable test conditions [6]. Thus, a panel of several listeners is often used for the scientific evaluation of speech; this is, however, quite time–consuming. Nevertheless, it is still the most commonly used method in scientific research to assess speech intelligibility [7, 8], phonematic disorders and temporal structure of speech [9, 10]. Until now, objective diagnostic tools for the assessment of speech intelligibility after treatment have only been performed for the quantification of nasalance [11], spectral characteristics, and the intensity of the voice signal [12]. Yet, these methods have limitations and do not allow assessing speech intelligibility in a comprehensive and reliable way.

A new technique for the objective evaluation of speech intelligibility has been used as a diagnostic tool in adult patients who suffered from neurological diseases [13], who stutter [14], for laryngectomees with tracheo–esophageal speech [15], and for children with cleft lip and palate [16, 17]. This method is based on a statistical analysis of speech with established methods of automatic speech recognition. It was the aim of the present study to test this method for the follow–up of patients treated for oral cancer and to compare the results of automatic evaluation of speech intelligibility with a perceptive rating of intelligibility by expert listeners.

2. Speech Data

In order to assess the speech of the patients, speech data was recorded using our **P**rogram for **E**valuation and **A**nalysis for all **K**inds of **S**peech (PEAKS) [18]. This software records speech data from an arbitrary client PC with Java Runtime Environment (JRE) 1.5.0.6 or higher. After recording the data is sent via an SSL encrypted connection to a server which is located at our university. Here all the analyses are performed. The evaluation result is available shortly after the recording at the client PC. All data are recorded at 16 kHz with 16 bit quantization using a close–talking microphone.

For this study we recorded 46 patients (13 female and 33 male) in the age of 34 to 80 (mean 60 $\pm$ 10). All of the patients were recorded after surgical treatment of the oral cavity. They read the German version of "The North Wind and the Sun", a fable from Aesop. It is a phonetically rich text with 108 words (71 disjoint).For the recording with PEAKS, the text was split into ten passages at major syntactic (i.e. sentence) boundaries in order to display the text in large letters which are well readable for elderly people without disturbing the reading flow. The recording software segments the audio data automatically according to these boundaries.

In order to get a reference for the intelligibility of the patients, four speech experts listened to the recordings and gave marks on a scale from 1 (very good) to 5 (very bad) for each turn and each patient. The final intelligibility score for each patient is obtained by averaging the marks of all turns and experts for each patient.

3. Automatic Speech Recognition System

In order to assess the intelligibility of the patients, we analyze the recognition rate of a word recognizer. In the following we describe how this value is computed.

A state–of–the–art word recognition system developed at the Chair of Pattern Recognition (Lehrstuhl für Mustererkennung) of the University of Erlangen–Nuremberg was used as described in detail in [19]. The recognizer can handle spontaneous speech with mid–sized vocabularies of up to 10,000 words. 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 [20]. 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 basic training set for our recognizer are dialogues from the VERBMOBIL project [21]. 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 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 [22, 23]). After the training, the vocabulary was reduced to the words occurring in the German version of the text "The North Wind and the Sun".

Several of the patients had difficulties in reading. Being unused to the situation they produced reading errors or asked the physician questions in the middle of the text. This results in additional words, which are not caused by recognition errors. Therefore, we computed the word recognition rate (WR) instead of the word accuracy (WA) to represent the intelligibility score. The recognition system, however, was optimized according to the WA during its training. The WR describes how many words of the text were correctly recognized in percent. It is calculated with the following formula:

$$\mathrm{WR} = C/R * 100\,\%$$

C is the number of correctly recognized words and R is the number of words of the reference text.

4. Analysis and Automatic Evaluation

To compute the agreement between different raters on the one hand and raters/recognizer on the other hand, we employed the Pearson's Product–Moment Correlation Coefficient. We calculated Spearman's Correlation Coefficient as well. The results only differed slightly. Therefore, we only mention Pearson's Correlation in this paper.

	1 rater	2 raters	3 raters
rater 1	.86 ± .03	.90 ± .01	.91
rater 2	.89 ± .07	.93 ± .03	.95
rater 3	.86 ± .09	.94 ± .03	.91
rater 4	.78 ± .03	.80 ± .02	.81
mean	.85 ± .07	.89 ± .06	.90 ± .07

Table 1: *Agreement of one human rater with an increasing number of reference raters: The more raters are used to create the gold standard the better is the agreement.*

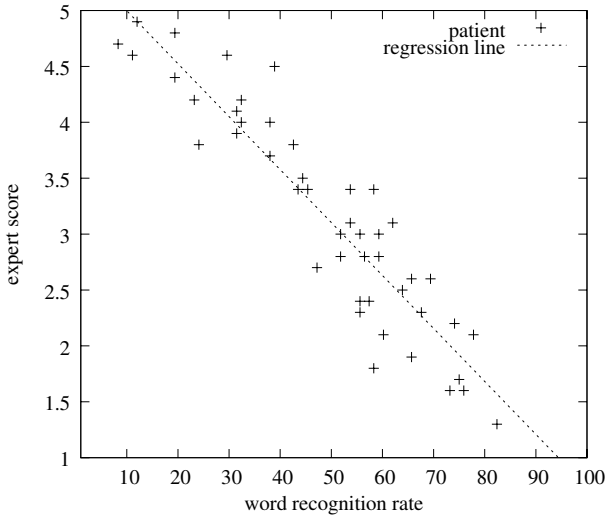

Figure 1: *The correlation between the word recognition rate (WR) and the intelligibility scores of the human experts is strong (r=-.92).*

5. Subjective Evaluation

A panel of four voice professionals subjectively estimated the intelligibility of the speech data while listening to a play–back of the recordings. 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 all individual turns. By that, an averaged mark — expressed as a floating point value — for each patient could be calculated.

Table 1 shows the agreement of the different raters. Note that the overall agreement between the raters increases, the more raters are used in the reference. We correlated each rater against all combinations of the other raters. For the case of three raters there is only one combination; for two and one rater as reference there are three possible combinations each since we have a total of four raters and one rater has to be excluded at a time. The mean is calculated as the average of all correlations of the corresponding number of reference raters. The multi–rater κ [24] for all raters was .55 (a κ value of .4 is considered as moderate agreement and a value of .75 as strong agreement).

6. Experimental Results

The WR is in the range between 8 % and 82 % (mean 49 % ± 19). Between experts' ratings and the automatic assessment (word recognition rate) exists a strong correlation (r= -.92, $p < .01$). Figure 1 shows the WR vs. the average intelligibility score of the four speech experts.

	1 rater	2 raters	3 raters	4 raters
WR	-.87 ± .03	-.91 ± .02	-.92 ± .01	**-.92**

Table 2: *Agreement between the automatic recognition system and the human raters: The more raters are used to create the reference the better is the agreement between the mean of the human raters and the automatic recognition system.*

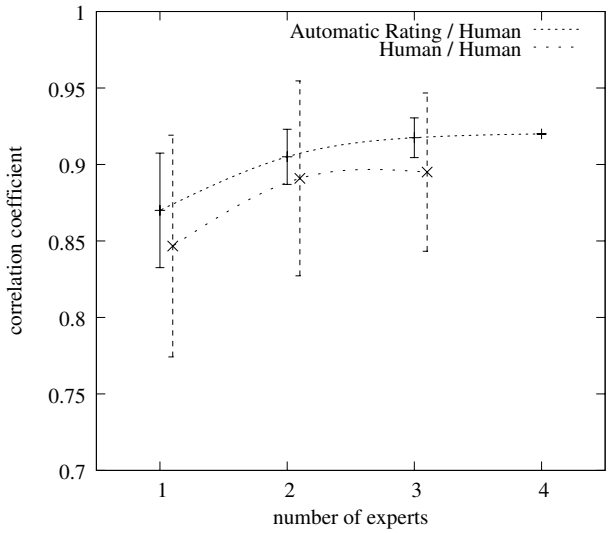

Figure 2: *Graphical visualization of the relation between the number of raters and the agreement: The more raters are used the higher is the agreement of the mean of these raters compared to a single rater. Note that the variance drops with the growing number of experts. The automatic system has the highest agreement to the mean of the experts and the lowest variance.*

An interesting observation is shown in Table 2: The more experts are used to create the reference (1 rater: 4 combinations; 2 raters: 6 combinations; 3 raters: 4 combinations; 4 raters: 1 combination) the better is the correlation between the mean of the experts and the automatic speech recognition rate. Figure 2 gives an even better impression of this relation. For the human experts as well as for the speech recognition system, the average correlation increases with a growing number of experts in the reference to which they are compared. On average the speech recognition has a better correlation to the mean of the experts than the human experts themselves. Furthermore, the speech recognition system shows less variance when compared to the human experts than the human experts have in their own group. We conclude, therefore, that the speech recognition system is more reliable than the individual speech expert. In addition, increasing the number of experts, whose average score is used as a reference, reduces the variance within the expert group; the same can be observed when the experts are compared to the automatic system. Thus this procedure is a good method for reducing the subjectivity of the "gold standard". In our experiments it is sufficient to use four experts since both the mean correlation and the variance converge to a stable value at this point.

7. Discussion

In the present study, a new method for the automatic evaluation of speech intelligibility is introduced. This technique analyses the word recognition rate (WR) of an automatic speech recognition system for a read standard text. The study revealed a relevant correlation between results of the automatic speech evaluation system and the experts' evaluation despite of the fact that evaluation of the speech intelligibility carried out by humans is hampered by a pronounced intra–individual variability. The limitations of speech evaluation by experts are highlighted by the results given in Table 1: although the experts' evaluations show a good correlation, they vary between different expert listeners. Such an imprecise assessment of speech intelligibility can be avoided by using an automatic speech evaluation that considers every single word and is independent of contextual information that influences perceptive ratings. Therefore, it describes the acoustic properties of speech more precisely and facilitates comparisons between different speech samples independently of time and place of recording. In general, speech recognition depends on five factors [25]:

- the speaker,
- the speech (read speech, spontaneous speech),
- the vocabulary,
- the grammatical complexity or perplexity (average probability of words possibly following a sequence of others),
- and the input medium.

For the diagnostic purpose, the influence of most of these factors was minimized by using a standard text and a stable setting. Thus, the speaker remains the main factor of influence.

Previously, automatic speech recognition techniques have been successfully used for the evaluation of communication disorders such as severe voice disorders of laryngectomees, stuttering, and speech disorders of children. The method showed a high correlation of the automatically evaluated intelligibility with perceptive ratings of a panel of experts. Now we demonstrated that the method can be applied to assessing speech disorders of adults. The correlation between four experts and the automatic evaluation of intelligibility is very high (r= -.92). To prove the reliability of the new method, patients with different extents of speech disorders as a consequence of the surgical therapy of oral cancer were examined. The disorders ranged from patients with small tumors whose speech was not disturbed up to patients with large tumors and severe speech disorders. It can be expected that the new method will be valuable and appropriate for clinical and scientific use. Further adaptation should enable to recognize different phonematic disorders. This will allow for comparisons of different surgical strategies concerning speech outcome and identify the appropriate but least impairing therapy strategy for oral cancer in the future.

It seems interesting to have a closer look at the subjective differences between the different raters. We found first evidence that these differences might be modeled by other features of speech. As found in [26] there are certain prosodic features which are correlated to the perception of intelligibility. Future investigations will identify prosodic influences on the variety of perceptive ratings.

8. Acknowledgments

This project was supported by ELAN Fonds of the University of Erlangen–Nuremberg.

9. References

[1] R. Schönweiler, C. Altenbernd, R. Schmelzeisen, and M. Ptok, "Articulatory capacity and intelligibility of speech of patients with carcinomas of the mouth cavity. a comparison of pre- and postoperative results of various reconstruction techniques," *HNO*, vol. 44, no. 11, pp. 634–639, 1996.

[2] J. Panchal, A. Potterron, E. Scanlon, and N. McLean, "An objective assessment of speech and swallowing following free flap reconstruction for oral cavity cancers," *British Journal of Plastic Surgery*, vol. 49, pp. 363–369, 1996.

[3] B. Pauloski, J. Logemann, L. Colangelo, A. Rademaker, F. McConnel, M. Heiser, S. Cardinale, D. Shedd, D. Stein, Q. Beery, J. Lewin, M. Haxer, and R. Esclamando, "Surgical variables affecting speech in treated patients with oral and oropharyngeal cancer," *Laryngoscope*, vol. 108, pp. 908–916, 1998.

[4] K. Mády, R. Sader, P. Hoole, A. Zimmermann, and H. Horch, "Speech evaluation and swallowing ability after intra-oral cancer," *Clinical Linguistic Phoniatrics*, vol. 17, pp. 411–420, 2003.

[5] P. Enderby, *Frenchay Dysarthrie Test*. Idstein, Germany: Schulz-Kirchner-Verlag, 1999.

[6] K. Keuning, G. Wieneke, and P. Dejonckere, "The intra-judge reliability of the perceptual rating of cleft palate speech before and after pharyngeal flap surgery: The effect of judges and speech samples," *Cleft Palate Craniofacial Journal*, vol. 36, no. 4, pp. 328–333, 1999.

[7] K. Robbins, J. Bowman, and R. Jacob, "Postglossectomy deglutitory and articulatory rehabilitation with palatal augmentation prostheses," *Archives Otolaryngology Head Neck Surgery*, vol. 113, pp. 1214–1218, 1987.

[8] R. Wachter and P. D. Dios, "Effect of adaptation and compensation mechanisms on postoperative function of patients with tumors of the oral cavity," *Laryngorhinootologie*, vol. 72, no. 7, pp. 333–337, 1993.

[9] G. Mahanna, D. Beukelman, J. Marshall, C. Gaebler, and M. Sullivan, "Obturator prostheses after cancer surgery: an approach to speech outcome assessment," *Journal of Prosthetic Dentistry*, vol. 79, pp. 310–316, 1998.

[10] B. Pauloski, A. Rademake, J. Logemann, and L. Colangelo, "Speech and swallowing in irradiated and nonirradiated postsurgical oral cancer patients," *Otolaryngology Head Neck Surgery*, vol. 118, pp. 616–624, 1998.

[11] C. Kuttner, R. Schönweiler, B. Seeberger, R. Dempf, J. Lisson, and M. Ptok, "Normal nasalance for the german language. nasometric values for clinical use in patients with cleft lip and palate," *HNO*, vol. 51, pp. 151–156, 2003.

[12] A. Zečević, "Ein sprachgestütztes Trainingssystem zur Evaluierung der Nasalität," Ph.D. dissertation, Universität Mannheim, Germany, 2002.

[13] B. Sy and D. Horowitz, "A statistical causal model for the assessment of dysarthric speech and the utility of computer-based speech recognition," *IEEE Transactions on Biomedical Engineering*, vol. 40, pp. 1282–1298, 1993.

[14] E. Nöth, H. Niemann, T. Haderlein, M. Decher, U. Eysholdt, F. Rosanowski, and T. Wittenberg, "Automatic Stuttering Recognition using Hidden Markov Models," in *Proceedings International Conference on Spoken Language Processing (ICSLP)*, vol. 4, Beijing, China, 2000, pp. 65–68.

[15] 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.

[16] 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.

[17] 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.

[18] A. Maier, E. Nöth, E. Nkenke, and M. Schuster, "Automatic Assessment of Children's Speech with Cleft Lip and Palate," in *Fifth Slovenian and First International Language Technologies Conference*, Ljubljana, Slovenia, 2006, pp. 31–35.

[19] G. Stemmer, *Modeling Variability in Speech Recognition*. Berlin: Logos Verlag, 2005.

[20] E. 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.

[21] W. Wahlster, *Verbmobil: Foundations of Speech-to-Speech Translation*. New York, Berlin: Springer, 2000.

[22] F. Gallwitz, *Integrated Stochastic Models for Spontaneous Speech Recognition*, ser. Studien zur Mustererkennung. Berlin, Germany: Logos Verlag, 2002, vol. 6.

[23] G. Stemmer, "Modeling Variability in Speech Recognition," Ph.D. dissertation, Chair for Pattern Recognition, University of Erlangen-Nuremberg, Germany, 2005.

[24] M. Davies and J. Fleiss, "Measuring agreement for multinomial data," *Biometrics*, vol. 38, pp. 1047–1051, 1982.

[25] F. Gallwitz, H. Niemann, and E. Nöth, "Speech recognition—state of the art, applications, and future prospects," *Wirtschaftsinformatik*, vol. 41, no. 6, pp. 538–547, 1999.

[26] 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: TUDpress, 2006, pp. 701–704.