Objective vs. Subjective Evaluation of Speakers with and without Complete Dentures

Tino Haderlein[1,2], Tobias Bocklet[1], Andreas Maier[1,2], Elmar Nöth[1], Christian Knipfer[3], and Florian Stelzle[3]

[1] Universität Erlangen-Nürnberg, Lehrstuhl für Mustererkennung (Informatik 5), Martensstraße 3, 91058 Erlangen, Germany
Tino.Haderlein@informatik.uni-erlangen.de
http://www5.informatik.uni-erlangen.de

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

[3] Universität Erlangen-Nürnberg, Mund-, Kiefer- und Gesichtschirurgische Klinik, Glückstraße 11, 91054 Erlangen, Germany

Abstract. For dento-oral rehabilitation of edentulous (toothless) patients, speech intelligibility is an important criterion. 28 persons read a standardized text once with and once without wearing complete dentures. Six experienced raters evaluated the intelligibility subjectively on a 5-point scale and the voice on the 4-point Roughness-Breathiness-Hoarseness (RBH) scales. Objective evaluation was performed by Support Vector Regression (SVR) on the word accuracy (WA) and word recognition rate (WR) of a speech recognition system, and a set of 95 word-based prosodic features. The word accuracy combined with selected prosodic features showed a correlation of up to $r = 0.65$ to the subjective ratings for patients with dentures and $r = 0.72$ for patients without dentures. For the RBH scales, however, the average correlation of the feature subsets to the subjective ratings for both types of recordings was $r < 0.4$.

1 Introduction

Complete loss of teeth can cause a persisting speech disorder by altering dental articulation areas. This reduces the intelligibility of speech severely. Removable complete dentures can partly solve this problem. However, they also disturb speech production as they restrict the flexibility of the tongue, narrow the oral cavity and alter the articulation areas of the palate and teeth.

Objective and independent diagnostic tools for the assessment of speech ability concerning alteration of the dental arch or dento-oral rehabilitation have only been applied for single parameters of speech. They do not evaluate continuous speech but often only single vowels or consonants [1–3]. However, this does not reflect real-life communication because no speech but only the voice is examined. Criteria like intelligibility cannot be evaluated in this way. For this study, the test persons read a given standard text which was then analyzed by methods of automatic speech recognition and prosodic analysis.

It was the aim of this clinical pilot study to evaluate speech intelligibility of edentulous patients objectively and automatically and to find out whether the impact of complete dentures on speech intelligibility can be evaluated by automatic analysis as part of oro-dental rehabilitation assessment.

In Sect. 2, the speech data used as the test set will be introduced. Section 3 will give some information about the speech recognizer. An overview on the prosodic analysis will be presented in Sect. 4, and Sect. 5 will discuss the results.

2 Test Data and Subjective Evaluation

The study group comprised 28 edentulous, i.e. toothless, patients (13 men, 15 women). Their average age was 64 years; the standard deviation was 10 years. The youngest person was 43, the oldest was 83 years old. They had worn their dentures on average for 59 months (st. dev. 49 months, range: 1 to 240 months) before the recordings. Only patients with removable complete dentures were accepted to participate to avoid the influence of different kinds of dentures. Only patients were chosen who wore them for more than at least one month to ensure patients' habituation to new dentures. All patients were native German speakers using the same local dialect. However, they were asked to speak standard German while being recorded. None of the patients had speech disorders caused by medical problems others than dental or any report of hearing impairment.

Each person read the text "Der Nordwind und die Sonne", a phonetically balanced text with 108 words (71 disjunctive) which is used in German speaking countries in speech therapy. The English version is known as "The North Wind and the Sun". The speech data were sampled with 16 kHz and an amplitude resolution of 16 bit. The patients read the text with their complete dentures inserted at first. The second recording was subsequently performed without dentures.

Six experienced phoniatricians and speech scientists evaluated each speaker's intelligibility in each recording according to a 5-point scale with the labels "very high", "high", "moderate", "low", and "none". Each rater's decision for each patient was converted to an integer number between 1 and 5. The 2·28 recordings were presented to the listeners during one evaluation session in random order.

Since the voice of elderly people is also often hoarse, the RBH scale [4] was applied which is an important rating system for dysphonic speech in German-speaking countries. It allows integer scores between 0 and 3 for the three dimensions "Roughness", "Breathiness", and "Hoarseness". The raters evaluated all recordings also with respect to these criteria. Since the RBH scales evaluate voice quality, it was expected that the dentures merely affect these ratings but rather the intelligibility scores.

3 The Speech Recognition System

The speech recognition system used for the experiments was developed at the Chair of Pattern Recognition in Erlangen [5]. It can handle spontaneous speech

with mid-sized vocabularies up to 10,000 words. The system is based on semi-continuous Hidden Markov Models (HMM). It can model phones in a context as large as statistically useful and thus forms the so-called polyphones, a generalization of the well-known bi- or triphones. The HMMs for each polyphone have three to four states; the codebook had 500 classes with full covariance matrices. The short-time analysis applies a Hamming window with a length of 16 ms, the frame rate is 10 ms. The filterbank for the Mel-spectrum consists of 25 triangle filters. For each frame, a 24-dimensional feature vector is computed. It contains short-time energy, 11 Mel-frequency cepstral coefficients, and the first-order derivatives of these 12 static features. The derivatives are approximated by the slope of a linear regression line over 5 consecutive frames (56 ms).

The baseline system for the experiments in this paper was trained with German dialogues from the VERBMOBIL project [6]. The data were recorded with a close-talking microphone at a sampling frequency of 16 kHz and quantized with 16 bit. About 80% of the 578 training speakers (304 male, 274 female) were between 20 and 29 years old, less than 10% were over 40. 11,714 utterances (257,810 words) of the VERBMOBIL-German data (12,030 utterances, 263,633 words, 27.7 hours of speech) were used for training and 48 (1042 words) for the validation set, i.e. the corpus partitions were the same as in [5].

The recognition vocabulary of the recognizer was changed to the 71 words of the standard text. The word accuracy and the word recognition rate were used as basic automatic measures for intelligibility since they had been successful for other voice and speech pathologies [7, 8]. They are computed from the comparison between the recognized word sequence and the reference text consisting of the $n_{\mathrm{all}} = 108$ words of the read text. With the number of words that were wrongly substituted (n_{sub}), deleted (n_{del}) and inserted (n_{ins}) by the recognizer, the word accuracy in percent is given as

$$\mathrm{WA} = [1 - (n_{\mathrm{sub}} + n_{\mathrm{del}} + n_{\mathrm{ins}})/n_{\mathrm{all}}] \cdot 100$$

while the word recognition rate omits the wrongly inserted words:

$$\mathrm{WR} = [1 - (n_{\mathrm{sub}} + n_{\mathrm{del}})/n_{\mathrm{all}}] \cdot 100$$

Only a unigram language model was used so that the results mainly depend on the acoustic models. A higher-order model would correct too many recognition errors and thus make WA and WR useless as measures for intelligibility.

4 Prosodic Features

In order to find automatically computable counterparts for the subjective rating criteria, also a "prosody module" was used to compute features based upon frequency, duration, and speech energy (intensity) measures. This is state-of-the-art in automatic speech analysis on normal voices [9–11].

The input to the prosody module is the speech signal and the output of the word recognition module. In this case the time-alignment of the recognizer

and the information about the underlying phoneme classes can be used by the module. For each speech unit which is of interest (here: words), a fixed reference point has to be chosen for the computation of the prosodic features. This point was chosen at the end of a word because the word is a well–defined unit in word recognition, it can be provided by any standard word recognizer, and because this point can be more easily defined than, for example, the middle of the syllable nucleus in word accent position. For each reference point, 95 features are extracted over intervals which contain a word, a word-pause-word interval or the pause between two words. A full description of the features used is beyond the scope of this paper; details and further references are given in [12]. The feature set was also used successfully for other voice and speech pathologies [7, 8].

In order to find the best subset of word accuracy, word recognition rate, and the prosodic features to model the subjective ratings, Support Vector Regression (SVR, [13]) was used. The general idea of regression is to use the vectors of a training set to approximate a function which tries to predict the target value of a given vector of the test set. Here, the training set were the automatically computed measures, and the test set consisted of the subjective intelligibility scores or the single dimensions of the RBH scores, respectively. For this study, the sequential minimal optimization algorithm (SMO, [13]) of the Weka toolbox [14] was applied in a 28-fold cross-validation manner due to the 28 available speakers.

5 Results and Discussion

The speech intelligibility of edentulous patients was rated lower by the speech experts on the 5-point scale (Table 1) where lower numbers denote better intelligibility. The average score was 2.19 for patients without teeth and 2.04 for the patients with dentures. The average RBH results, their range and standard deviation for both cases were virtually identical, but a closer analysis revealed that this does not hold for each single patient (Fig. 1). Pearson's correlation r was computed for each rater against the average of the other 5 raters and then averaged (Table 2). For the intelligibility criterion, an average of $r = 0.73$ (edentulous patients) and $r = 0.74$ (with dentures) was reached. The values did not change significantly throughout the study when Spearman's rank-order correlation ρ was computed. For this reason, only r will be given in the following. For the RBH scales, they were in the same range for the R and H scale but only when the patients did not wear their dentures. The reason for this may be the rough 4-point RBH scales that do not allow a better differentiation in rating.

The automatically computed word accuracy and word recognition rate were also lower for the edentulous patients (WA: 55.8%, WR: 63.1%; see Table 1) than for the same persons with complete dentures (WA: 59.4%, WR: 68.2%). The ranges of both measures were shifted by about the same value as the averages. The correlation between subjective evaluation and WA or WR, respectively, was lower than among the rater group (Table 3). For the average rater's intelligibility scores, the WA reached $r = -0.53$ for patients without and $r = -0.60$ for patients with complete denture. The corresponding values for the WR were $r = -0.55$ and

$r = -0.46$. The coefficient is negative because high recognition rates came from "good" voices with a low score number and vice versa.

The RBH scores could not achieve satisfying correlation with WA or WR (Table 3). The best correlation was $r = -0.50$ for the WR on patients without teeth. When the patients wore their dentures, the human-machine correlation dropped drastically although there was a high correlation both for the subjective and the objective evaluations when the two recordings of each patient were compared (Table 4). Obviously, WA and WR are not suitable to reflect slight changes in signal quality that the trained listeners can hear.

By using WA, WR, and the prosodic features as input for SVR, higher correlations to the subjective intelligibility score were achieved (Table 5). For edentulous patients $r = 0.72$ was reached when the WA value was combined with the normalized energy computed from the current word and the energy from the two words before the current word and the pause between them. The F_0 value at the end of the last voiced section within a respective word contributes also to these results. For patients with complete denture, $r = 0.65$ was reached. In that case, however, the energy value from the current word was not beneficial. In general, the number of speakers in this study was rather small and the results have to be handled with care. However, the contribution of the mentioned features to the human-machine correlation was evident throughout the experiments.

In order to explain the influence of the speech energy, it would be straightforward to assume that a louder speaker is better intelligible. However, in the prosodic feature set, the energy values are normalized so that a continuously high energy level will have no effect. It is more likely that single phones or phone classes that cannot be uttered properly due to the speech impairment appear in the signal as more noisy and cause local changes in the energy distribution.

The impact of the F_0 value can be explained by the noisy speech that causes octave errors during F_0 detection, i.e. instead of the real fundamental frequency, one of its harmonics one or more octaves higher is found. Again, with more "noisy speech", this may influence the F_0 trajectory and hence the correlation to the subjective results. It is not clear so far, however, why only the end of the voiced sections causes a noticeable effect. There may be a connection to changes in the airstream between the beginning and the end of words or phrases, but this has to be confirmed by more detailed experiments. An aspect that also needs a closer look in the future is that not all phone clases are affected in the same way by missing teeth. Especially the articulation of fricatives, like /s/, is distorted. The word-based analysis will therefore be extended by a phone-based level.

For the RBH scores, the SVR method could not reveal a feature set that showed good results on both types of recordings so far. The average correlations were below $r = 0.40$. Further experiments will be part of future work.

For this study, patients read a standard text, and voice professionals evaluated intelligibility. It is often argued that intelligibility should be evaluated by an "inverse intelligibility test": The patient utters a subset of words and sentences from a carefully built corpus. A naïve listener writes down what he or she heard. The percentage of correctly understood words is a measure for the

Table 1. Subjective and objective evaluation results for 28 speakers: intelligibility (int.), roughness (R), breathiness (B), hoarseness (H), all of them averaged across 6 raters, and the word accuracy (WA) and word recognition rate (WR) in percent

	without denture						with denture					
	int.	R	B	H	WA	WR	int.	R	B	H	WA	WR
mean	2.19	0.89	0.31	0.96	55.8	63.1	2.04	0.90	0.32	0.93	59.4	68.2
st. dev.	0.65	0.56	0.29	0.54	12.4	8.8	0.68	0.48	0.28	0.47	15.0	8.6
min.	1.17	0.00	0.00	0.00	13.9	46.3	1.17	0.00	0.00	0.00	5.6	51.9
max.	3.67	2.17	1.17	2.17	75.0	79.6	3.83	2.17	1.00	2.17	85.2	86.1

Table 2. Average inter-rater correlation for intelligibility (int.), roughness (R), breathiness (B), and hoarseness (H)

	without denture				with denture			
	int.	R	B	H	int.	R	B	H
r	0.73	0.73	0.38	0.71	0.74	0.57	0.31	0.56

intelligibility of the patient. However, when automatic speech evaluation is performed for instance with respect to prosodic phenomena, like e.g. word durations or percentage of voiced segments [7], then comparable results for all patients can only be achieved when all the patients read the same defined words or text. This means that an inverse intelligibility test can no longer be performed, and intelligibility has to be rated on a grading scale instead.

The data obtained in this study allow for the following conclusions: There is a significant correlation between subjective rating of intelligibility and automatic evaluation. It also reveals the impact of dentures on intelligibility just as the subjective ratings do. Hence, the method can serve as the basis for more research towards an automatic system that can support oro-dental rehabilitation by objective speech evaluation.

Acknowledgments

This study was partially funded by the German Cancer Aid (Deutsche Krebshilfe) under grant 107873. The responsibility for the contents lies with the authors.

References

1. Runte, C., Tawana, D., Dirksen, D., Runte, B., Lamprecht-Dinnesen, A., Bollmann, F., Seifert, E., Danesh, G.: Spectral analysis of /s/ sound with changing angulation of the maxillary central incisors. Int J Prosthodont **15** (2002) 254–258
2. Molly, L., Nackaerts, O., Vandewiele, K., Manders, E., van Steenberghe, D., Jacobs, R.: Speech adaptation after treatment of full edentulism through immediate-loaded implant protocols. Clin Oral Implants Res **19** (2008) 86–90

Table 3. Correlation between subjective ratings for intelligibility (int.), roughness (R), breathiness (B), and hoarseness (H) vs. the word accuracy (WA) and word recognition rate (WR) for patients without denture and with full denture, respectively

	without denture				with denture			
	int.	R	B	H	int.	R	B	H
r(WA)	–0.53	–0.42	–0.40	–0.43	–0.60	–0.20	0.04	–0.23
r(WR)	–0.55	–0.49	–0.35	–0.50	–0.46	–0.14	0.01	–0.20

Table 4. Correlation between the ratings for patients without denture and with full denture: intelligibility (int.), roughness (R), breathiness (B), hoarseness (H), word accuracy (WA), and word recognition rate (WR)

	subjective				objective	
	int.	R	B	H	WA	WR
r	0.64	0.80	0.55	0.77	0.80	0.72

3. Stojcević, I., Carek, A., Buković, D., Hedjever, M.: Influence of the partial denture on the articulation of dental and postalveolar sounds. Coll Antropol **28** (2004) 799–807
4. Nawka, T., Anders, L.-C., Wendler, J.: Die auditive Beurteilung heiserer Stimmen nach dem RBH-System. Sprache - Stimme - Gehör **18** (1994) 130–133
5. Stemmer, G.: Modeling Variability in Speech Recognition. Volume 19 of Studien zur Mustererkennung. Logos Verlag, Berlin (2005)
6. Wahlster, W., ed.: Verbmobil: Foundations of Speech-to-Speech Translation. Springer, Berlin (2000)
7. Haderlein, T.: Automatic Evaluation of Tracheoesophageal Substitute Voices. Volume 25 of Studien zur Mustererkennung. Logos Verlag, Berlin (2007)
8. Maier, A.: Speech of Children with Cleft Lip and Palate: Automatic Assessment. Volume 29 of Studien zur Mustererkennung. Logos Verlag, Berlin (2009)
9. Nöth, E., Batliner, A., Kießling, A., Kompe, R., Niemann, H.: VERBMOBIL: The Use of Prosody in the Linguistic Components of a Speech Understanding System. IEEE Trans. on Speech and Audio Processing **8** (2000) 519–532
10. Chen, K., Hasegawa-Johnson, M., Cohen, A., Borys, S., Kim, S.S., Cole, J., Choi, J.Y.: Prosody dependent speech recognition on radio news corpus of American English. IEEE Trans. Audio, Speech, and Language Processing **14** (2006) 232–245
11. Shriberg, E., Stolcke, A.: Direct Modeling of Prosody: An Overview of Applications in Automatic Speech Processing. In: Proc. International Conference on Speech Prosody, Nara, Japan (2004) 575–582
12. Batliner, A., Buckow, A., Niemann, H., Nöth, E., Warnke, V.: The Prosody Module. [6] 106–121
13. Smola, A., Schölkopf, B.: A Tutorial on Support Vector Regression. Statistics and Computing **14** (2004) 199–222
14. Witten, I., Frank, E., Trigg, L., Hall, M., Holmes, G., Cunningham, S.: Weka: Practical machine learning tools and techniques with java implementations. In: Proc. ICONIP/ANZIIS/ANNES99 Future Directions for Intelligent Systems and Information Sciences, San Francisco, Morgan Kaufmann (1999) 192–196

Table 5. Correlation of subjective intelligibility scores from 6 raters against a feature subset computed by Support Vector Regression (SVR) from the values of word accuracy (WA), word recognition rate (WR), and 95 prosodic features

	without denture	with denture
r(subset with energy of current word)	0.52	0.72
r(subset without energy of current word)	0.65	0.63

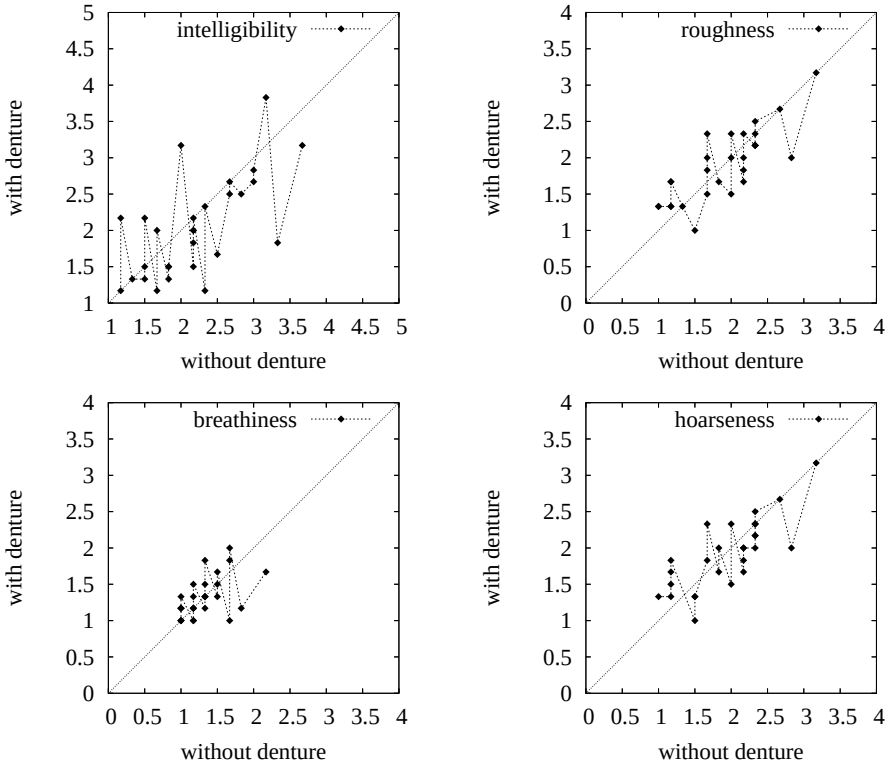

Fig. 1. Subjective rating criteria for the patients with and without dentures, respectively. Note that all measures are ordered independently from each other.