Automatic Quantification of Speech Intelligibility of Adults with Oral Squamous Cell Carcinoma

M. Windrich  A. Maier  R. Kohler  E. Nöth  E. Nkenke  U. Eysholdt  M. Schuster

Department of Phoniatrics and Pediatric Audiology, Universitätsklinikum Erlangen, Erlangen, Germany

Introduction

Oral squamous cell carcinoma is one of the ten most common malignant diseases [1]. Its treatment often negatively affects expressive speech skills and reduces speech intelligibility. As a consequence, patients’ postoperative quality of life can be significantly impaired [2]. There is no standardized assessment of speech disorders in adults or children at national or international level [3–5]. Moreover, a method that allows for assessing speech disorders objectively and independently is missing. Therefore, it is difficult to compare the outcome of different therapeutic options in a reliable way.

Semistandardized instruments for the analysis of speech disorders in adults are well known [6–10]. But the assessment of speech disorders or intelligibility is usually performed subjectively and therefore lacks reliability due to differences in individuals’ experience and variable test conditions [11]. Therefore, a panel of several listeners is often used for scientific evaluation of speech. This method is still the most widely used technique for assessing speech intelligibility [12–20], phonematic disorders and temporal structure of speech [21–26]. For more reliable results, transcription tasks and multiple-choice tasks for several listeners have been found to be appropriate [7, 9, 15, 17, 19, 21, 25]. Unfortunately, the use of several listeners is rather time-consuming and is thus only used for...
research projects. For clinical purposes, usually only one expert evaluates the patient’s speech.

Objective and independent diagnostic tools for the assessment of speech intelligibility after treatment have only been performed for the quantification of nasalance [27] and the spectral characteristics and intensity of the voice signal [28]. However, these methods do not allow for assessing speech intelligibility in a comprehensive and reliable way. A new technique for objective evaluation of speech intelligibility has been tested as a diagnostic tool in adult patients who suffered from neurological diseases [29], who stuttered [30, 31], in laryngectomees with tracheo-esophageal speech [32] and in children with cleft lip and palate [33, 34]. 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 squamous cell carcinoma and to compare the results of automatic evaluation of speech intelligibility with a perceptual rating of intelligibility by a panel of expert listeners.

Material and Methods

Patients

The study group comprised 46 patients with different speech disorders. Each patient was recorded during regular outpatient care (Table 1). All of them had suffered from squamous cell carcinoma of the mouth (Table 2). They were treated by surgery, microvascular reconstruction of the defect for tumors with at least T2 staging, underwent modified radical neck dissection on the ipsilateral side and suprathyroid neck dissection on the contralateral side and had adjuvant radiation therapy. All patients were native German speakers, using a local dialect. Forty subjects without oral diseases speaking the same local dialect served as the control group (Table 1).

The patients read a German text (‘Der Nordwind und die Sonne’, ‘The North Wind and the Sun’), a fable by Aesop. It is known as the reference text for the International Phonetic Alphabet by the International Phonetic Association. The German text consists of 108 words. It is phonetically balanced and includes all possible phonemes of the German language. The text was divided into 10 sequences (10.8 ± 2.4 words) according to syntactic boundaries and shown on a computer screen. The speech samples were recorded with a close-talking microphone (Call4U Comfort-Headset, DNT GmbH, Dietzenbach, Germany, sampling frequency 16 kHz, 16 bit).

Automatic Speech Recognition System

Intelligibility was measured by means of a state-of-the-art automatic speech recognition system [34]. It segments speech data into temporal units of 16 ms. Then temporal and spectral characteristics are analyzed and compared with word models. A speech recognition system evaluates two different information channels: acoustic signals modeled by stochastic word models (hidden Markov model) and word frequency in the text represented by a stochastic language model. A speech recognition system typically has a so-called bi- or trigram-language model. Hence, the probability of a word depends on the acoustic signal (obtained by the hidden Markov models) and the probability that this word follows the last spoken (bigram) or the two last spoken words (trigram). For our purposes we used a unigram-language model, which assumes that the current word is independent of previously spoken words. The only linguistic parameter included is the frequency of each word in the recognition vocabulary. Thus, recognition mainly depends on the acoustic signal of each single word. Our recognition system is polyphone-based at the acoustic level, i.e. the acoustic properties of a phoneme are computed with respect to the coarticulatory modulation caused by its phonetic context. It is well known [35] that the realization of a phoneme is influenced by its phonetic context. This is why most speech recognition systems use triphone models as elementary recognition units, i.e. a different model is trained for each phonetic context. For instance, a /ʃ/ /t/ /ɹ/ model for /t/ with /ʃ/ on the left and /ɹ/ on the right is trained with exactly this context. Word models then concatenate the respective elementary models. In our system, we use a pho-

<table>
<thead>
<tr>
<th>Table 1. Age and gender of 46 patients with oral squamous cell carcinoma and the control group of 40 subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Range</td>
</tr>
<tr>
<td>Mean ± SD</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2. Clinical TNM classification of patient group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Patients</td>
</tr>
</tbody>
</table>
netic context of arbitrary length (polyphones), provided it is observed often enough in the training database (more than 50 times). The acoustic model has been trained with speech recordings of 578 adult speakers (304 male, 274 female) without any speech disorder.

We computed the so-called word recognition rate (WR) of the speech data using the PEAKS evaluation software. The computation time is less than half of real time. The WR describes the percentage of correctly recognized words in the whole text. It is calculated as follows:

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

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

**Perceptual Evaluation of Speech Intelligibility**

Three experienced clinicians and a speech recognition expert, who was well trained also in perceptual evaluation, estimated speech intelligibility independently by listening to a playback of the recordings presented via headphones. None of the expert listeners was familiar with the patients. The recordings were presented segment by segment and judged after each segment. The listeners used a five-point Likert scale (1 = very high, 2 = rather high, 3 = medium, 4 = rather low, 5 = low) to rate the intelligibility of all segments of the recordings. The average of these estimations formed the final score on a quasi-continuous scale for every record.

**Statistics**

The statistical analysis was performed using Microsoft Excel and our own evaluation software implemented in Java, which computed Pearson’s product-moment correlation coefficient (r), Spearman’s rank correlation coefficient (ρ), Student’s t test, and multirater kappa [37]. The latter poses a special problem: the averages of listener ratings and word accuracy are continuous measures with completely different ranges. Thus, correlation techniques should be applied. If one wants to use kappa, mapping onto discrete values has to be performed. Therefore, we rounded the listeners’ average intelligibility scores to the next integer and set thresholds on the WR results, so that the difference between the listeners’ scores and the scores derived from the WR values was minimal. Although thresholds should not be set on the test data, the small size of our database does not allow us to define a validation set. Therefore, the kappa values for the WR results should be interpreted with caution.

**Results**

**Subjective Rating**

Listener ratings showed good agreement (table 3). The multirater kappa for all raters was 0.55. (A kappa value of 0.4 is considered as moderate agreement and a value of 0.75 as strong agreement.) Pearson’s correlation for all raters (r = 0.90 ± 0.06) was found to be very close to Spearman’s correlation (ρ = 0.89 ± 0.06).

**Comparison between Subjective and Objective Assessment of Intelligibility**

There was a strong correlation between listeners’ ratings and automatic assessment (word recognition rate) (r = −0.93; ρ = −0.90; p < 0.01, multirater kappa 0.58). The

<table>
<thead>
<tr>
<th>Listener</th>
<th>Listener</th>
<th>Listener</th>
<th>Listener</th>
<th>Mean of all other listeners</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.82</td>
<td>0.87</td>
<td>0.90</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.74</td>
<td>0.80</td>
<td>0.81</td>
</tr>
<tr>
<td>1</td>
<td>0.96</td>
<td>0.91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>1</td>
<td>0.95</td>
<td></td>
</tr>
</tbody>
</table>

‘All’ (right column) means the average of the remaining 3 listeners that is compared to each single listener.

![Fig. 1. Comparison of the listeners’ ratings of intelligibility and WR. Perceptual rating is performed using categories from 1 (very high intelligibility) to 5 (very low intelligibility). WR is the percentage of correctly recognized words of the standard text. There is a high correlation between perceptually judged intelligibility scores and the automatic recognition system (r = −0.93).](image-url)
distance between objective evaluation and subjective rating was less than one point on the Likert scale (fig. 1).

**Comparison between Patients and Control Group**

Automatically computed data of the control group showed significantly better WR values than for the patients (table 4). The WR of the control group was ranging from 60 to 91%. There was no correlation between age and WR in the patient group (r = –0.06; p > 0.3), while a weak correlation in the control group (r = –0.3; p < 0.05) was found.

**Discussion**

In the present study, a new method for the automatic evaluation of speech intelligibility is introduced. This technique is based on the WR of spoken language as a means of representing intelligibility by speech recognition technique. The study revealed a significant correlation between the results of the automatic speech evaluation system and listeners’ evaluation, despite the fact that evaluation of speech intelligibility by listeners is always biased due to considerable intraindividual variability.

The limitations of speech evaluation by listeners are evident in the results given in table 3. Although the listeners’ evaluations show good correlation, they vary between listeners. Several methods exist for perceptual assessment. We decided to use a five-point Likert scale because it produced similar results as visual analogue scales in preliminary experiments. Furthermore, quasi-linear scales are well-tested tools for voice evaluations, e.g. GRBAS or RBH scales. Nevertheless, Whitehill [4] and Schiavetti [38] questioned the validity of equidistant scales. Schiavetti [38] pointed out that ‘listeners cannot effect a linear partition of the dimension with the result that the intervals used to scale speech intelligibility are not equal’.

Nevertheless, he adds that practical limitations may force one to use a suboptimal method. So we decided for a descriptive five-point scale. To analyze the informative value of the Likert scale for this application, we compared Pearson’s correlation coefficient with Spearman’s rank correlation, since the latter is independent of the underlying scale used in an experiment. Spearman’s correlation considers only the rank of the ratings instead of the actual value on the underlying scale. As described in the ‘Results’ section, both results were almost identical. For our experiments we used a standard test for all patients for better comparability, e.g. when comparing different patients or one patient’s intelligibility over time. This, however, precludes the use of transcription or multiple-choice tasks as applied by other authors [39, 40]. Furthermore, transcriptions by multiple naive listeners have comparable inter-rater correlation as perceptual evaluations [41].

In general, speech recognition depends on five factors: the speaker, speech (read speech, spontaneous speech), vocabulary, grammatical complexity or perplexity (average probability of words possibly following a sequence of others), and the input medium [42]. For diagnostic purposes, 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 as required for the evaluation of intelligibility [38].

Automatic speech recognition techniques have been used successfully in the evaluation of communication disorders, such as severe voice disorders in laryngectomies, stuttering and speech disorders in children. In these cases the method showed a close correlation of automatically evaluated intelligibility with perceptual ratings by a panel of expert listeners. The present study shows that this method can also be used in speech disorders of adults. The correlation between 4 expert listeners and the automatic evaluation of intelligibility is very high (r = –0.93). To demonstrate the reliability of the new method, patients with a wide range of speech disorders formed the cohort. The study group included patients with small tumors, whose speech was only minimally disturbed, but also patients with large tumors and severe speech disorders. When speech intelligibility was perceived as high or very high, WR was within the range of the control group. Comparison between the data of the control group and those of the study group showed that speech intelligibility was significantly reduced (p < 0.01) in the study group (table 4). Moreover, the results of the control group demonstrated the variability of ‘normal’ speech that also depends on age. The finding that some

<table>
<thead>
<tr>
<th>WR, %</th>
<th>Range</th>
<th>Mean ± SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients’ speech (n = 46)</td>
<td>8–82</td>
<td>49 ± 19</td>
</tr>
<tr>
<td>Speech of control group (n = 40)</td>
<td>60–91</td>
<td>76 ± 7</td>
</tr>
</tbody>
</table>
of the speakers from the control group had poor results is not surprising. The recognizer was trained with mostly young people. In comparison, elderly people and children deviate from the ‘standard’ speaker and within the speaker group [43], showing lower WR. Also, the speech recognition system mostly uses acoustic information (‘unigram language model’), which leads to restricted word recognition results. One has to keep in mind that absolute WR is not crucial, it is dependent on the training population and adaptations of the system, i.e. a speech recognition system that achieves a better recognition rate on the same corpus does not necessarily reflect intelligibility more accurately. The recognizer should rather cover the whole spectrum of intelligibility levels. Such is the case with the applied system, where the results vary from 8 to 91% and the range of the patient group differs from the range of the control group: \( (\mu \pm \sigma)_{\text{patient group}} < (\mu - \sigma)_{\text{control group}} \).

In the future, a larger control group will allow for correcting patient data for age and gender effects on the WR in order to quantify deviations of a patient’s WR from normal speech. It can be expected that the new method will be valuable and appropriate for clinical and scientific use. Automatic speech evaluation considers every single word and is independent of prosodic, pragmatic and contextual information that influence perceptual ratings. Therefore, it describes the acoustic properties of speech precisely and facilitates comparison between different speech samples independent of time and place of recording.

Further adaptations should enable the recognition of different phonetic disorders. In the future this will enhance the scientific evaluation of different surgical strategies concerning speech outcome and the identification of the best and least impairing treatment modality for oral cancer. Nevertheless, as with every diagnostic tool it does not replace the expert’s interpretation and detailed examination.

**Conclusion**

The automatic speech recognition system is a valuable technique for investigating speech intelligibility of adult patients treated for oral squamous cell carcinoma. It objectifies and quantifies global speech outcome for clinical and scientific purposes. The method is easily available and applicable for the global assessment of speech outcome.

**Acknowledgments**

This work was supported by the ELAN-Fonds of Friedrich-Alexander University of Erlangen-Nuremberg. We thank Dr. Ralf Frisch for his valuable comments on this article.

---

**References**


16 Brown JS, Zuydam AC, Jones DC, Rogers SN, and reconstruction with the radial forearm free flap. Laryngoscope 2003;113:897–904.


