Hindawi Publishing Corporation
EURASIP Journal on Audio, Speech, and Music Processing
Volume 2010, Article ID 926951, 7 pages
doi:10.1155/2010/926951

Research Article

Automatic Speech Recognition Systems for the Evaluation of Voice and Speech Disorders in Head and Neck Cancer

Andreas Maier,[1] Tino Haderlein,[1] Florian Stelzle,[2] Elmar Nöth,[3] Emeka Nkenke,[2] Frank Rosanowski,[1] Anne Schützenberger,[1] and Maria Schuster[1]

[1] *Division of Phoniatrics and Pediatric Audiology, Erlangen University Hospital, Friedrich-Alexander University Erlangen-Nuremberg, Bohlenplatz 21, 91054 Erlangen, Germany*
[2] *Department of Maxillofacial Surgery, Erlangen University Hospital, Friedrich-Alexander University Erlangen-Nuremberg, Glückstraße 11, 91054 Erlangen, Germany*
[3] *Department of Computer Science, Friedrich-Alexander University Erlangen-Nuremberg, Martensstraße 3, 91058 Erlangen, Germany*

Correspondence should be addressed to Andreas Maier, andreas.maier@cs.fau.de

Received 6 November 2008; Revised 9 April 2009; Accepted 16 June 2009

Academic Editor: Georg Stemmer

In patients suffering from head and neck cancer, speech intelligibility is often restricted. For assessment and outcome measurements, automatic speech recognition systems have previously been shown to be appropriate for objective and quick evaluation of intelligibility. In this study we investigate the applicability of the method to speech disorders caused by head and neck cancer. Intelligibility was quantified by speech recognition on recordings of a standard text read by 41 German laryngectomized patients with cancer of the larynx or hypopharynx and 49 German patients who had suffered from oral cancer. The speech recognition provides the percentage of correctly recognized words of a sequence, that is, the word recognition rate. Automatic evaluation was compared to perceptual ratings by a panel of experts and to an age-matched control group. Both patient groups showed significantly lower word recognition rates than the control group. Automatic speech recognition yielded word recognition rates which complied with experts' evaluation of intelligibility on a significant level. Automatic speech recognition serves as a good means with low effort to objectify and quantify the most important aspect of pathologic speech—the intelligibility. The system was successfully applied to voice and speech disorders.

1. Introduction

Head and neck cancer may affect speech and voice. In many cases, surgical treatment will even deteriorate the patients' ability for oral communication and thus affect their quality of life [1]. So, rehabilitation of dysglossia and/or dysphonia is of outstanding clinical interest. However, assessment of these functional features often lacks objectivity, as it is mainly performed by expert rating which is time- and manpower-consuming and—although being the gold standard in the clinical field—questionable for scientific purpose. Even speech pathologists with high expertise reach only low reliability when judging disturbed speech [2]. Therefore, a panel of several listeners may be used for evaluation of speech disorders and temporal structure of speech [3]. Of course, this makes assessment even more time-consuming, "expensive," and inadequate for clinical application. So, there is need for automatic evaluation of speech and voice.

However, objective assessment of speech disorders and severe voice disorders are neither nationally nor internationally standardized [4]. For a long time, automatic diagnostic tools for quantitative assessment of speech and voice are restricted to single aspects such as the quantification of nasalance in text passages, spectral characteristics, and intensity of the voice signal in sustained vowels [5]. Moreover, most commonly used methods have limitations for severely disordered voice or speech and do not allow for assessing speech intelligibility in a comprehensive and reliable way.

The use of speech processing methods for speech intelligibility assessment is getting more and more popular [6].

Van Nuffelen et al. presented an automatic measure for the phoneme intelligibility using phonological features in dysarthric Dutch speech [7]. As perceptual reference the Dutch Intelligibility Assessment (DIA) was applied to the speech data to measure the phoneme intelligibility. In the dysarthric speech data, they obtain correlations between the perceptual evaluation and their automatic measure of up to 0.94. By now the procedure was also extended to be applicable to other types of disorders [8].

Visualization of speech and voice disorders helps the medical staff in the assessment of voice and speech disorders. The Sammon transform is suitable to reduce the dimensionality of the speech data to create a map which allows the comparison between different speakers [9]. Also the influence of the recording conditions can be reduced efficiently [10].

In this study, we describe a new automatic method for the evaluation of speech intelligibility in comparison to traditional perceptual evaluation. The new method is based on automatic speech recognition techniques and was tested on speech of patients with different but significant extents of speech disorder (dysglossia) and voice disorder (dysphonia) after the treatment for head and neck cancer.

2. Speech Disorders in Patients with Oral Cancer

Dysglossia often occurs in patients with oral squamous cell carcinoma which belongs to the ten most frequent malignant diseases [11]. The extent of dysglossia depends on many factors such as the stage of the tumour, its localisation, but also on the individual treatment procedure.

The main feature in speech of patients with oral cancer is compensatory articulation [12]. Often the movement capability of the tongue is restricted. Hence, the patients have to compensate this by alternative speech gestures. While the primary voice signal, that is, the fundamental frequency, is in a normal range, the formants and their range may be affected. Sumita et al. showed that the first formant (F1) is higher in the vowel /i/, and the second formant (F2) is lower in vowels /a/, /e/, /i/, /o/, and /u/ compared to a control group [13]. The range of F2 was also shown to be significantly lower in patients with oral cancer.

Hence, the speech impairment basically consists of reduced articulation skills affecting not only consonants and consonant clusters [14], but also vowels [13]. This leads to reduced intelligibility.

3. Voice Disorders in Laryngectomees

For the evaluation of extensive voice disorders, we chose a group of patients with severe dysphonia after total laryngectomy due to laryngeal or hypopharyngeal cancer. All patients used tracheo-esophageal substitute voice, which is regarded as the state of the art for voice rehabilitation [14]. After removal of the larynx, the breathing ability is maintained by a hole in the neck. A one-way shunt valve is placed between the trachea and the oesophagus. Then the patient can create an artificial substitute voice by breathing in and closing the hole in the neck. When the patient breathes out, the air will be detoured through the one-way valve and stream from the trachea into the oesophagus. The tissue in the oesophagus will start to oscillate and create the so-called substitute voice.

Although the substitute voice resembles laryngeal voice production more than alternative techniques [15], it still shows considerable differences to laryngeal voices. It is characterized by high perturbation causing roughness of the voice and reduced prosody. It shows low fundamental frequency, short maximum phonation time, and a different ratio of voiced to voiceless phonation in comparison with normal speech. All these aspects lead to significantly decreased intelligibility.

Hence, in both disorders, intelligibility is the superordinate functional outcome parameter, and so, the present study focuses on this essential feature.

4. Speech Data

All patients and control subjects read the text "Der Nordwind und die Sonne," a tale from Aesop known as "The North Wind and the Sun" in the Anglo-American world. It is a phonetically rich text with 108 words (71 disjunctive) often used in speech assessment in German-speaking countries. It is also used as reference text for the International Phonetic Alphabet by the International Phonetic Association (IPA). The text was divided into 10 sequences (11 ± 2.4 words each), according to syntactic boundaries, and shown on a computer screen. The speech samples were recorded with a close-talk microphone (Call4U Comfort Headset, DNT GmbH, Dietzenbach, Germany; sampling frequency 16 kHz, amplitude resolution 16 bit).

The dysglossia study cohort (OC) comprised 49 patients (14 females and 35 males) after treatment of oral squamous cell carcinoma, graded T1 to T4. The patients' mean age was 60.1 ± 10.4 years. The treatment included the excision of the tumour and additional radiotherapy for most of the patient except for T1 grading. Recording was performed at least 3 months after the treatment had been completed.

The dysphonia patient group (LE) consisted of 39-male and 2-female laryngectomees. Their average age was 62.0 ± 7.7 years. They had undergone total laryngectomy because of T3 or T4 laryngeal or hypopharyngeal cancer at least one year prior to the investigation and were provided with a Provox shunt valve for tracheo-esophageal substitute speech.

At the time of the investigation, none of the patients suffered from recurrent tumour growth or metastases. All patients had been informed about the scientific character of the study and had given their informed consent.

From each patient acoustic data were recorded during regular out patient care. All patients were native German speakers using the same local dialect.

40 subjects (10 females and 30 males) without oral or laryngeal diseases or malignoma of any kind speaking the same local dialect formed the control group (CON). The control group was age matched (58.1 ± 13.3 years old) with respect to the patient groups.

5. Perceptual Evaluation

A panel of voice professionals perceptually evaluated the intelligibility of each patient while listening to a play back of the recordings. A five-point Likert scale was applied to rate the intelligibility of all individual samples (1 = "very high," 2 = "rather high," 3 = "medium," 4 = "rather low," 5 = "very low"). For the LE group, five raters were asked to use the total range from 1 to 5 and to set 1 for "very good substitute speech" instead of "very good normal speech." For the four raters of the OC patients there was no need to alter the Likert scale.

For both databases the mean score of all perceptual evaluations was computed for each patient. This expert mean score was used to represent the patient's speech intelligibility.

6. Automatic Speech Recognition System

For objective measurement, an automatic speech recognition (ASR) system was applied. We use an automatic speech recognition system based on Hidden Markov Models (HMMs). The word recognition system was developed at the Chair of Pattern Recognition at the University of Erlangen-Nuremberg. In this study, the version as described in detail by Stemmer [16] was used. A commercial version of this ASR system is used for conversational dialogue systems (http://www.sympalog.com/).

As features we use 11 Mel-Frequency Cepstrum Coefficients (MFCCs) and the energy of the signal plus their first derivatives. The short-time analysis applies a Hamming window with a length of 16 milliseconds. The frame rate is 100 Hz. The filter bank for the Mel-spectrum consists of 25 triangular filters. Delta coefficients of the 12 static features are computed over a context of 2 time frames to the left and to the right side (56 milliseconds in total).

Our recognition system works polyphone based on the acoustic level, that is, the acoustic attributes of a phoneme are computed with respect to the coarticulatory modulation caused by its phonetic context, for example, the pronunciation of "i" in "bird" is different than the "o" in "worth" although both realize the same phoneme /3:/. The pronunciation depends on the phonetic context. Sometimes this includes more than only the neighbouring phonemes. The construction of polyphones is data driven according to the number of observed phoneme sequences in the training set, that is, if a context appears more than 50 times in the training data then a polyphone is constructed. The HMMs for the polyphones have three to four states.

An ASR system normally has a so-called bi- or trigram language model. For our purpose we used several language models to investigate the dependency between the recognition performance and the correlation to the experts' perceptual evaluation. With the ASR system, we calculated the word recognition rate (WR) of the recordings [17]:

$$\mathrm{WR}\ [\%] = \frac{C}{R} * 100\%. \tag{1}$$

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

Table 1: Effect of the n-gram language model on the oral cancer (OC) data: with growing context of the language model the recognition rate increases. Correlation ρ to the perceptual evaluation, however, decreases if the context is too large (starting with $n = 3$ here). Higher n-grams showed even worse performance.

ρ	0-gram	1-gram	2-gram	3-gram
Correlation	−0.88	−0.90	−0.90	−0.85
WR in %	44.4	50.0	67.3	74.8

The basic training sets for our recognizer are dialogues from the Verbmobil project [18]. 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 males, 274 females) 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 was more than 60 years may influence the recognition results.

A subset of the German Verbmobil data was used for the training set (11,714 utterances, 257,810 words, 27 hours of speech) and 48 utterances (1042 words) for the validation set (the training and validation corpora were the same as in [16]).

This ASR system is integrated into the "Program for the Evaluation and Analysis of all Kinds of Speech disorders" (PEAKS) which can be used via the Internet [19]. Results are available shortly after the recordings. The actual processing speed is faster than real time. All data is transmitted encrypted in order to guarantee the security of the data. Furthermore, the patient data are pseudonymized. Hence, even if the encryption was broken, no personal information could be obtained by nonauthorized persons.

Statistical analysis was performed using SPSS [20]. For the agreement computations between different raters on the one hand, and raters/speech recognition system on the other hand, Spearman's correlation coefficient was used. Contrary to other agreement measurements, such as Kappa and Alpha, correlations are suitable to compare the averaged scores of the raters and the WR even though both scales differ in their order of magnitude. Comparisons of the mean values were performed using Student's T-test. The test for normal distribution of an input variable was performed using the Kolmogorov-Smirnov test.

7. Results

The recordings showed a wide range in intelligibility. The perceptual evaluation resulted in 2.9 ± 1.0 for the oral cancer group on the five-point scale and 2.5 ± 1.0 for the group of laryngectomees.

The effect of the language model is investigated in Table 1 using the OC data. With growing n-gram context, the recognition rate increases (cf. Figure 1). However, the increased n-gram context is not always beneficial for the

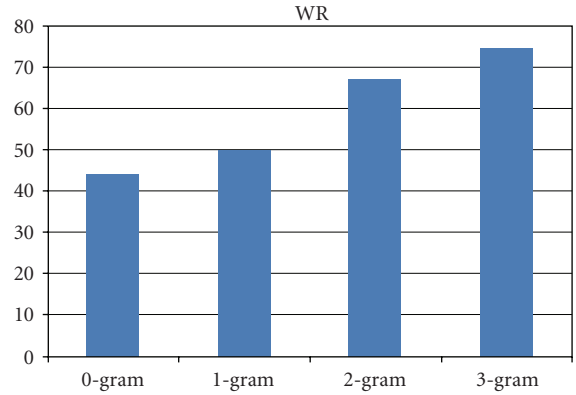

Figure 1: Effect of the language model context on the recognition rate: with growing context the recognition rate increases (OC data).

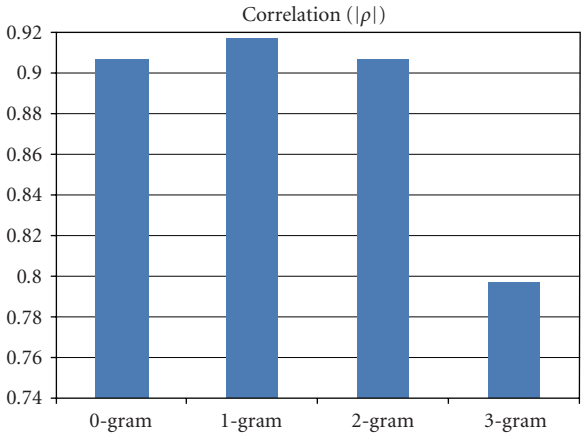

Figure 2: Effect of the language model context on the correlation to the perceptual evaluation: if the language model context is increased too much, the absolute value of the correlation to the perceptual evaluation decreases (here starting at $n = 3$). Higher n-gram contexts showed even worse performance.

correlation to the perceptual evaluation. Figure 2 shows this effect. While the correlations when using zero- to bigram language models lie in the same range, the correlation using the trigram model is reduced. With higher n-gram context, even more severe reductions were observed. Hence, we decided for a unigram language model in the following experiments.

Table 2 shows the results of the automatic speech recognition system. In both patient groups, a high variance in the ASR results was found. In the group with oral cancer (OC), WR (49 ± 19) differs significantly from the control group's WR ($P < .001$) with $76 \pm 7\%$. Also compared to the LE group (48 ± 14), the recognition results were significantly higher in the control group ($P < .001$).

Inter-rater correlations were computed for each of the raters using the respective other raters as reference, that is, the mean of the four other raters for the LE group and the three other raters for the OC group. For the comparison between the automatic speech recognition system and the expert ratings, the mean rating of all human raters was taken into account. Table 3 shows the results of the correlation analysis. The inter-rater correlations of the experts' ratings range from 0.76 to 0.88. The correlations between the automatic system and the experts' ratings are negative since a high expert rating represents a low intelligibility and hence a low recognition rate and vice versa. The agreement between WR and the mean scores of the perceptual ratings is very high in both patient groups and is in the same range as the best human experts with −0.83 and −0.90, respectively. The experts' ratings differ, compared to the automatic rating, less than one point with respect to the regression line (see Figures 3 and 4).

Table 2: Results of the automatic recognition of the speech recordings: the table presents the percentage of correctly recognized words of a sequence (WR) read by laryngectomees (LE), patients with oral cancer (OC) and a control group (CON).

	WR min.; max. in %	WR mean ± standard deviation in %
OC	8; 82	50 ± 19
LE	17; 74	48 ± 14
CON	60; 91	76 ± 7

Table 3: Spearman's rank correlations ρ of each individual rater versus the mean value of the other raters and the mean of the raters versus the speech recognition system. The correlation is negative for the latter because the scales for the evaluations are in opposite directions.

	OC	LE
rater 1	0.84	0.82
rater 2	0.76	0.84
rater 3	0.88	0.77
rater 4	0.82	0.83
rater 5	—	0.77
WR	−0.90	−0.83

8. Discussion

During and after therapy of malignant diseases of the head and neck, communication skills are of special interest. Until now, no objective method with low effort and costs existed to determine speech and voice outcome. Here, we present a new automatic objective measurement of speech quality based on automatic speech recognition (ASR). It quantifies speech intelligibility as good as the former clinical standard procedure. According to common recommendations for diagnostics of voice and speech disorders [5], the voice and speech function can now be objectified and quantified. The new method might close the gap between the exact description of morphologic impairment by endoscopic and imaging methods and the standardized perceptual evaluation of the individual handicap. ASR will enable precise evaluations of speech and voice as a precondition for scientific purposes, for example, for outcome measurements. It will help to specify the influence of therapy options such as different surgical procedures and nonsurgical therapies on communication

skills [21] and the role of speech and voice on the patients' experiences [22].

For the perceptual quantitative assessment, two different kinds of scales are widely used [23]: equal-appearing interval scales and direct magnitude estimation scales. In recent literature it has been discussed which type of scale is more suitable for perceptual assessment. Some authors decide for direct magnitude estimation scales such as visual analogue scales, because the interval size of an equal-appearing interval scale might be not equal in all continua. Hence, statistical analyses which do not take this fact into account might be problematic. We still decided in favour of the equal-appearing interval scale for the sake of comparability with earlier results of our group. Hence, for statistical analyses one has to keep in mind that the equal-appearing intervals might not be equal. Therefore, also distances and errors which are usually optimized in such regression problems in a least-square error sense might not be equal. Hence, the Spearman rank order correlation [20] is appropriate. It is based on the idea that the input data might not be equally partitioned. Thus, the distances between the data points are declared meaningless. Instead, their rank, that is, their sequence, is considered to hold the important information. In this manner also equal-appearing scales such as Likert scales are reliably examined in a statistical regression analysis. Other agreement measures, for example, Cohen's Kappa or Alpha, were not used in this work since they are only applicable if all scales are defined in the same margin.

The limitations of perceptual speech evaluation by experts are seen when the results of the different experts, given in Table 3, are compared to each other. Although the experts' evaluations show a good correlation, their mean scores vary considerably between different expert listeners [24]. Previous studies even yielded more variability for linear ratings of laryngectomees' speech [25]. These findings support the need for objective and automatic speech evaluation.

We examined speech samples from laryngectomized speakers and patients after treatment for oral cancer. As a precondition for the evaluation of the patient groups, the extent of disorder is widespread, as seen by the results of the perceptual assessment and the ASR results. The presented method might also be applicable to all kinds of patient groups with other voice, speech, and language disorder, such as dysarthria and aphasia, since the subordinate parameter of speech—its intelligibility—is evaluated with the method [26]. Investigation of these disorders seems to be beneficial in the future. Even analysis of the emotional state of a patient is within the reach of ASR methods [27]. However, this is not the scope of the present work.

Today, ASR is used in many domains [28]: for professional and private use as dictating machines, in call centres when a restricted vocabulary, and "normal" voice quality and speech without background noise can be expected, and in the support of handicapped persons. Normally, ASR is meant to recognize speech as good as possible, and the technique that analyses speech signals and calculates the most probable word sequence is more and more refined. We

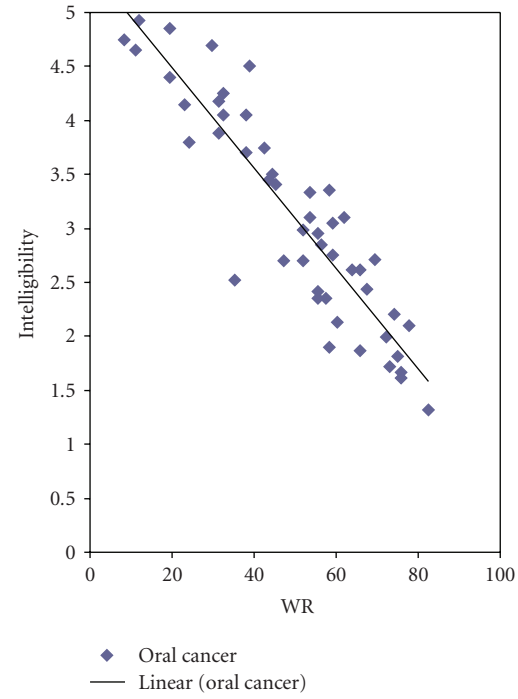

Figure 3: Comparison between the automatic evaluation, that is, the word recognition rate (WR) and the perceptual intelligibility evaluation by 5 expert listeners on the dysphonia (LE) database ($\rho = -0.83$).

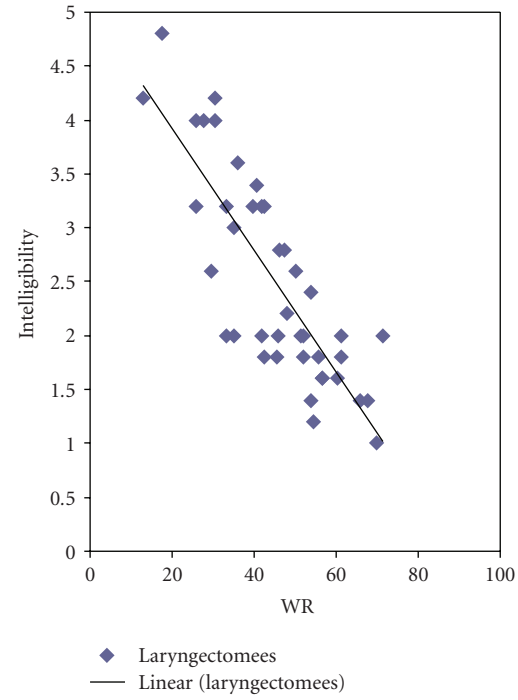

Figure 4: Comparison between the automatic evaluation, that is, the word recognition rate (WR) and the perceptual intelligibility evaluation by 4 expert listeners on the dysglossia (OC) database ($\rho = -0.90$).

use the technique for diagnostics to quantify the influence of altered speech and voice on the recognition results in stable ASR conditions. The quality of the recognition allows assessing the quality of the speech signal. In order to exclude methodical interferences, a standard text and a stable recording setup are used. Thus, the speaker remains the only factor of influence.

For this study, we applied a nonadapted ASR system for automatic speech evaluation that has previously been proven to be adequate for "normal" speech samples [16]. The automatic speech evaluations were compared to a control group of 40 speakers without speech pathology in this study. As increased age has been shown to have a negative influence on automatic speech recognition [29], the control group consisted of speakers of similar age compared to the patient groups. In our study, control speakers reached a word recognition rate of 76 ± 7% on average. This result might seem relatively low compared to other applications of ASR. Here it is caused by the use of a unigram language model which excludes semantic or contextual information. Therewith, the speech recognition is mostly based on acoustic properties of the words. Further use of linguistic knowledge indeed improves the recognition rate of the system as shown with the OC data, but the improvement by language modelling diminishes the impact of articulation and voice on the WR. In order to compare the perceptual evaluation with the automatic one, the differences in the recognition rates are more important than the absolute values, that is, the recognition rate does not need to be 100% as shown on the OC data (cf. Figure 2).

The result of our procedure is the percentage of correctly recognized words of a sequence—the word recognition rate. It is robust to reading errors. If a reader repeats or corrects himself, only the correct word is regarded for the evaluation since the system ignores additional wrong words. Hence, a low WR represents a reduced percentage of correctly recognized words. This corresponds to the perception by a human listener and, therefore, reflects the definition of intelligibility.

Until now, the technique has only been tested for the German language but might also be appropriate for other languages. A transfer to other speech or voice disorders is possible and has yet been shown for children's speech [30]. For clinical purpose the technique is embedded into a recording and evaluation software that can be easily accessed via the Internet on any PC provided with a microphone and a sound card [17]. The computation of the WR is quickly performed in less than real time, this is what makes the method a time- and manpower-saving procedure. Hence, the method is suitable for everyday clinical use.

We prefer ASR over other speech evaluation techniques based on forced alignment (FA) which is often used in second language learning. FA would allow only for one reference speaker. In this case it would be unclear who would be the best reference speaker. It is also unclear how FA should deal with self-corrections of the speaker. This would be regarded as an error in the FA. Of course, one could use all control speakers and their variations for the FA and compute some kind of mean score. However, this would be computationally much more expensive. The proposed method is supposed to work in real time.

The results of the control group demonstrated that the standard deviation in WR of "normal" speech in speakers of the same age is about half of the pathologic one. This is still considerable. Currently, norm data for all age classes and gender are not available. These could quantify a patient's intelligibility in relation to the norm in percent ranks. In the future, by using a larger control group, we will be able to provide age- and gender-dependent values for the WR. Then the deviation from normal speech will even be quantified exactly for each patient.

For the clinician, our novel method will allow an easy-to-apply, automated observer-independent evaluation of all kinds of voice and speech disorders in less than real time via the Internet.

9. Conclusion

Speech evaluation by an automatic speech recognition system is a valuable means for research and clinical purpose in order to determine the global functional outcome of speech and voice after the treatment of head and neck cancer. It allows quantifying the intelligibility also in severely disturbed voices and speech.

Acknowledgments

This work was supported by the Deutsche Krebshilfe (Grant no. 106266) and the ELAN-Fonds of the University Erlangen-Nuremberg. The authors are responsible for the content of this article.

References

[1] N.-C. Gellrich, R. Schimming, A. Schramm, D. Schmalohr, A. Bremerich, and J. Kugler, "Pain, function, and psychologic outcome before, during, and after intraoral tumor resection," *Journal of Oral and Maxillofacial Surgery*, vol. 60, no. 7, pp. 772–777, 2002.

[2] S. Paal, U. Reulbach, K. Strobel-Schwarthoff, E. Nkenke, and M. Schuster, "Beurteilung von Sprechauffälligkeiten bei Kindern mit Lippen-Kiefer-Gaumen-Spaltbildungen," *Journal of Orofacial Orthopedics*, vol. 66, no. 4, pp. 270–278, 2005.

[3] T. Bressmann, R. Sader, T. L. Whitehill, and N. Samman, "Consonant intelligibility and tongue motility in patients with partial glossectomy," *Journal of Oral and Maxillofacial Surgery*, vol. 62, no. 3, pp. 298–303, 2004.

[4] H. P. Zenner, "The postlaryngectomy telephone intelligibility test (PLTT)," in *Speech Restoration via Voice Prosthesis*, I. F. Herrmann, Ed., pp. 148–152, Springer, Berlin, Germany, 1986.

[5] P. H. Dejonckere, P. Bradley, P. Clemente, et al., "A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. Guideline elaborated by the Committee on Phoniatrics of the European Laryngological Society (ELS)," *European Archives of Oto-Rhino-Laryngology*, vol. 258, no. 2, pp. 77–82, 2001.

[6] P. Kitzing, A. Maier, and Å. Lyberg, "Automatic speech recognition (ASR) and its use as a tool for assessment or

therapy of voice, speech, and language disorders," *Logopedics Phoniatrics Vocology*, vol. 34, no. 2, pp. 91–96, 2009.

[7] G. Van Nuffelen, C. Middag, M. De Bodt, and J. P. Martens, "Speech technology-based assessment of phoneme intelligibility in dysarthria," *International Journal of Language and Communication Disorders*, vol. 30, pp. 1–15, 2008.

[8] C. Middag, J. P. Martens, G. Van Nuffelen, and M. De Bodt, "Automated intelligibility assessment of pathological speech using phonological features," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, Article ID 629030, 9 pages, 2009.

[9] T. Haderlein, D. Zorn, D. Steidl, E. Nöth, M. Shozakai, and M. Schuster, "Visualization of voice disorders using the sammon transform," in *Proceedings of the Text, Speech and Dialogue*, P. Sojka, I. Kopecek, and K. Pala, Eds., pp. 589–596, Springer, 2006.

[10] A. Maier, M. Schuster, U. Eysholdt, et al., "QMOS—a robust visualization method for speaker dependencies with different microphones," *Journal of Pattern Recognition Research*, vol. 4, no. 1, pp. 32–51, 2009.

[11] World Health Organization, "The World Oral Health Report 2003," World Health Organization, Geneva, Switzerland, 2003.

[12] D. A. Georgian, J. A. Logemann, and H. B. Fisher, "Compensatory articulation patterns of a surgically treated oral cancer patient," *Journal of Speech and Hearing Disorders*, vol. 47, no. 2, pp. 154–159, 1982.

[13] Y. I. Sumita, S. Ozawa, H. Mukohyama, T. Ueno, T. Ohyama, and H. Taniguchi, "Digital acoustic analysis of five vowels in maxillectomy patients," *Journal of Oral Rehabilitation*, vol. 29, no. 7, pp. 649–656, 2002.

[14] D. H. Brown, F. J. M. Hilgers, J. C. Irish, and A. J. M. Balm, "Postlaryngectomy voice rehabilitation: state of the art at the millennium," *World Journal of Surgery*, vol. 27, no. 7, pp. 824–831, 2003.

[15] J. Robbins, H. B. Fisher, E. C. Blom, and M. I. Singer, "A comparative acoustic study of normal, esophageal, and tracheoesophageal speech production," *Journal of Speech and Hearing Disorders*, vol. 49, no. 2, pp. 202–210, 1984.

[16] G. Stemmer, *Modelling Variability in Speech Recognition*, Logos, Berlin, Germany, 2005.

[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] W. Wahlster, Ed., *Verbmobil: Foundations of Speech-to-Speech Translation*, Springer, Berlin, Germany, 2000.

[19] A. Maier, T. Haderlein, U. Eysholdt, et al., "PEAKS—a system for the automatic evaluation of voice and speech disorders," *Speech Communication*, vol. 51, no. 5, pp. 425–437, 2009.

[20] R. Levesque, *SPSS Programming and Data Management: A Guide for SPSS and SAS Users*, SPSS, Chicago, Ill, USA, 4th edition, 2007.

[21] Scottish Intercollegiate Guidelines Network (SIGN), Diagnosis and management of head and neck cancer. A national clinical guideline, Edinburgh (Scotland): Scottish Intercollegiate Guidelines Network (SIGN), SIGN publication; no. 90, October 2006.

[22] K. MacKenzie, A. Millar, J. A. Wilson, C. Sellars, and I. J. Deary, "Is voice therapy an effective treatment for dysphonia? A randomised controlled trial," *British Medical Journal*, vol. 323, no. 7314, pp. 658–661, 2001.

[23] N. Schiavetti, "Scaling procedures for the measurement of speech intelligibility," in *Intelligibility in Speech Disorders: Theory, Measurement and Management*, R. D. Kent, Ed., pp. 11–34, John Benjamins, Philadelphia, Pa, USA, month 1992.

[24] M. Windrich, A. Maier, R. Kohler, et al., "Automatic quantification of speech intelligibility of adults with oral squamous cell carcinoma," *Folia Phoniatrica et Logopaedica*, vol. 60, no. 3, pp. 151–156, 2008.

[25] M. Schuster, T. Haderlein, E. Nöth, J. Lohscheller, U. Eysholdt, and F. Rosanowski, "Intelligibility of laryngectomees' substitute speech: automatic speech recognition and subjective rating," *European Archives of Oto-Rhino-Laryngology*, vol. 263, no. 2, pp. 188–193, 2006.

[26] R. J. Ruben, "Redefining the survival of the fittest: communication disorders in the 21st century," *Laryngoscope*, vol. 110, no. 2, part 1, pp. 241–245, 2000.

[27] A. Ozdas, R. G. Shiavi, D. M. Wilkes, M. K. Silverman, and S. E. Silverman, "Analysis of vocal tract characteristics for near-term suicidal risk assessment," *Methods of Information in Medicine*, vol. 43, no. 1, pp. 36–38, 2004.

[28] C. M. Karat, J. Vergo, and D. Nahamoo, "Conversational interface technologies, IBM Research," in *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications*, A. Sears and J. A. Jacko, Eds., Human Factors and Ergonomics, Lawrence Erlbaum Associates, 2007.

[29] J. G. Wilpon and C. N. Jacobsen, "Study of speech recognition for children and the elderly," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '96)*, vol. 1, pp. 349–352, 1996.

[30] M. Schuster, A. Maier, T. Haderlein, et al., "Evaluation of speech intelligibility for children with cleft lip and palate by means of automatic speech recognition," *International Journal of Pediatric Otorhinolaryngology*, vol. 70, no. 10, pp. 1741–1747, 2006.