Influence of Reading Errors on the Text-Based Automatic Evaluation of Pathologic Voices

Tino Haderlein[1,2], Elmar Nöth[1], Andreas Maier[1,2], Maria Schuster[2], and Frank Rosanowski[2]

[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

Abstract. In speech therapy and rehabilitation, a patient's voice has to be evaluated by the therapist. Established methods for objective, automatic evaluation analyze only recordings of sustained vowels. However, an isolated vowel does not reflect a real communication situation. In this paper, a speech recognition system and a prosody module are used to analyze a text that was read out by the patients. The correlation between the perceptive evaluation of speech intelligibility by five medical experts and measures like word accuracy (WA), word recognition rate (WR), and prosodic features was examined. The focus was on the influence of reading errors on this correlation.
The test speakers were 85 persons suffering from cancer in the larynx. 65 of them had undergone partial laryngectomy, i.e. partial removal of the larynx. The correlation between the human intelligibility ratings on a five-point scale and the machine was $r = -0.61$ for WA, $r \approx 0.55$ for WR, and $r \approx 0.60$ for prosodic features based on word duration and energy. The reading errors did not have a significant influence on the results. Hence, no special preprocessing of the audio files is necessary.

1 Introduction

Although less than 1% of all cancers affect the larynx, it is necessary to provide proper rehabilitation therapies since speech is the main means of communication. In the USA, 10,000 new cases of laryngeal cancer are diagnosed each year [1]. In severe cases total laryngectomy has to be performed, i.e. the removal of the entire larynx. In early and intermediate stages, usually partial laryngectomy is sufficient, and at least one of the vocal folds or the vestibular folds can be preserved (see Fig. 1). Dependent on the location and size of the tumor, the voice may sound normal before and after surgery. However, hoarse voices are very common.

In speech therapy and rehabilitation, a patient's voice has to be evaluated by the therapist. Automatically computed, objective measures are a very helpful support for this task. Established methods for objective evaluation, however, analyze only recordings of sustained vowels in order to find irregularities in the voice (see e.g. [2, 3]). However, this does not reflect a real communication situation 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. A standard text was used especially in view of the prosodic evaluation because the comparability of results among the patients is reduced when the utterances differ with respect to duration, number of words, percentage of different phone classes, etc.

For speech after total laryngectomy, where the patients use a substitute voice produced in the upper esophagus, we showed in previous work that an automatic speech recognition system can be used to rate intelligibility [4]. The word accuracy of the speech recognizer was identified as suitable measure for this task. It showed a correlation of more than $|r| = 0.8$ to the human evaluation. However, these results relied on the assumption that the recognition errors were only caused by the acoustic properties of the voices. Another source of error are reading errors. When the recognized word sequence is compared to the text reference, a patient with a high-quality voice might get bad automatic evaluation results due to misread words. This problem could be solved by replacing the text reference by a transliteration of the respective speech sample, but this method is not applicable in clinical practice.

In this paper, we examined how severe the influence of reading errors is on the results of automatic evaluation and the correlation to human evaluation results. In Sect. 2, the speech data used as the test set will be introduced. Section 3 will give some information about the speech recognition system. An overview on the prosodic analysis will be presented in Sect. 4. Section 5 contains the results, and Sect. 6 will give a short outlook on future work.

2 Test Data

The test files were recorded from 85 patients (75 men, 10 women) suffering from cancer in different regions of the larynx. 65 of them had already undergone partial laryngectomy, 20 speakers were still awaiting surgery. The former group was recorded on the average 2.4 months after surgery. The average age of all speakers was 60.7 years with a standard deviation of 9.7 years. The youngest and the oldest person were 34 and 83 years old, respectively.

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" [5]. The speech data were sampled with 16 kHz and an amplitude resolution of 16 bit.

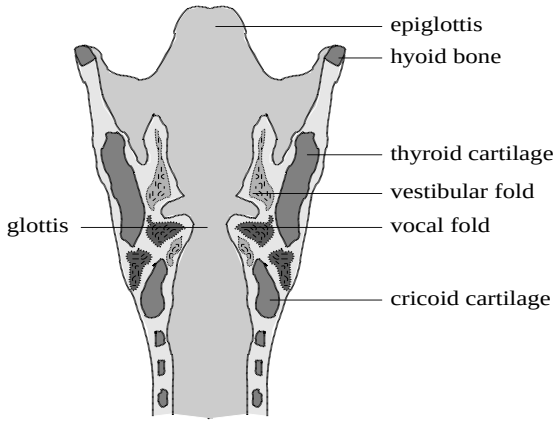

Fig. 1. Anatomy of an intact larynx

Table 1. File statistics for the speech corpora with and without reading errors

	duration					words	vocabulary
	total	avg.	st. dev.	min.	max.		
with errors	89 min	63 s	18 s	43 s	144 s	9519	71+187
without errors	82 min	58 s	15 s	40 s	125 s	9151	71+83

In order to obtain a reference for the automatic evaluation, five experienced phoniatricians and speech scientists evaluated each speaker's intelligibility 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.

Due to reading errors, repetitions and remarks like "I don't have my glasses with me.", the vocabulary in the recordings did not only contain the 71 words of the text reference but also 187 additional words and word fragments. 27 of the files were error-free. In all other samples, at least one error occurred (see Fig. 2). In order to determine the influence of these phenomena on the evaluation results, a second version of the data set was created by removing the additional words where possible. In total, 368 (3.9%) of the 9519 words were eliminated from the original speech samples.

Since the text flow was supposed to be preserved, misreading of single words without corrections, i.e. word substitutions, were not removed. This means that for instance the correction "*M*or- Nordwind" was reduced to "Nordwind" while the word "Nordw*u*nd" without correction was left unchanged. Also breaks in words, like "gel- -ten" were not changed when the full word was not repeated. This explains why the corrected files, further denoted as "without errors", still contain 83 out-of-text words and word fragments (see Table 1).

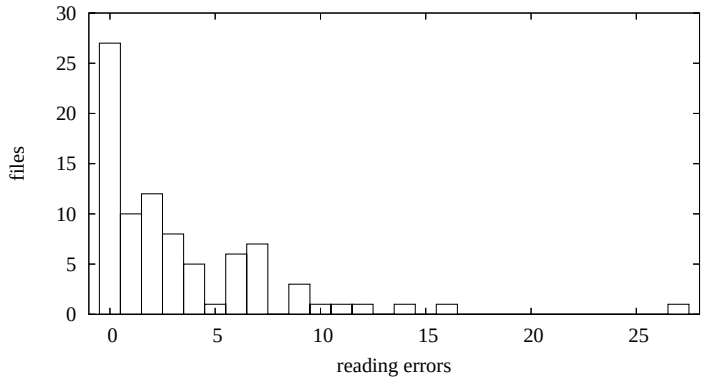

Fig. 2. Absolute number of reading errors in the 85 text samples

3 The Speech Recognition System

The speech recognition system used for the experiments was developed at the Chair of Pattern Recognition in Erlangen [6]. 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). A unigram language model was used so that the results are mainly dependent on the acoustic models.

The baseline system for the experiments in this paper was trained with German dialogues from the VERBMOBIL project [7]. The topic in these recordings is appointment scheduling. 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 the training and 48 (1042 words) for the validation set, i.e. the corpus partitions were the same as in [6].

The recognition vocabulary of the recognizer was changed to the 71 words of the standard text. The uttered word fragments and out-of-text words were

not added to the vocabulary because in a clinical application it will also not be possible to add the current reader's errors to the vocabulary in real-time.

4 Prosodic Features

In order to find automatically computable counterparts for subjective rating criteria, we also use a "prosody module" 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 [8–10].

The prosody module takes the output of our word recognition module in addition to the speech signal as input. In this case the time-alignment of the recognizer and the information about the underlying phoneme classes (like *long vowel*) can be used by the prosody 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. We decided in favor of 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, we extract 95 prosodic features over intervals which contain one single 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 [11].

Besides the 95 local features per word, 15 global features were computed per utterance from jitter, shimmer and the number of voiced/unvoiced (V/UV) decisions. They cover each of mean and standard deviation for jitter and shimmer, the number, length and maximum length each for voiced and unvoiced sections, the ratio of the numbers of voiced and unvoiced sections, the ratio of length of voiced sections to the length of the signal and the same for unvoiced sections. The last global feature is the standard deviation of the fundamental frequency F_0.

We examined the prosodic features of our speech data because for substitute voices after total laryngectomy we had found that several duration-based features showed correlations of up to $|r| = 0.72$ between human and automatic evaluation of intelligibility [12, p. 117]. The agreement was measured as the correlation between the mean value of the respective feature in a recording and the average expert's intelligibility rating for that file.

5 Results

The absolute recognition rates for the pathologic speakers when using a unigram language model were at about 50% for word accuracy (WA) and word recognition rate (WR; see Table 2). This was expected since the speech recognizer was trained with normal speech because the amount of pathologic speech data was too small for training. On the other hand, the recognizer simulates a "naïve" listener that has never heard pathologic speech before. This represents the situation that speech patients are confronted with in their daily life.

The average WA and WR rose only non-significantly when the reading errors were removed from the audio files. However, in some cases the recognition results got slightly worse when reading errors – mainly at the beginning of the files – were removed (see Fig. 3). The benefit of these sections for channel adaptation were obviously higher than the harm caused by the out-of-text utterances.

The agreement among the human raters when judging the speakers' intelligibility was $r = 0.81$. This value was computed as the mean of all correlations obtained when one of the raters was compared to the average of the remaining four. The correlation between the average human and the automatic evaluation for all 85 speakers was about $r = -0.6$ (see Table 2). The coefficient is negative because high recognition rates came from "good" voices with a low score number and vice versa. There is no significant difference in the correlation for the speech data with and without reading errors.

The human-machine correlation was also computed for the subgroup of speakers whose WA and WR were better after error correction and for the subgroup of the remaining speakers. 50 speakers showed improved WA; the correlation was $r = -0.67$ both for the files with and without errors. The other 35 patients reached $r = -0.52$ before and $r = -0.51$ after elimination of the reading errors. 49 speakers showed improved WR in the repaired files; the correlation dropped slightly from $r = -0.54$ to $r = -0.53$. The other 36 patients reached $r = -0.59$ in the original files and $r = -0.57$ in the files without errors.

It was expected that for recordings with a lot of reading errors the word recognition rate would achieve higher correlation to the human rating. This was based on the assumption that human raters are not affected in their judging of intelligibility when a speaker utters words that are not in the text reference. However, the correlation for WR was in all experiments of this study smaller than for the word accuracy.

For the agreement between human evaluation and prosodic features, the findings of [12] were confirmed. The same prosodic features as for substitute voices showed the highest correlation also for the 85 speakers of this study (see Table 3). The duration and pause-based features are highly correlated to the human intelligibility criterion since non-fluent pathologic speakers often show low voice quality and hence low intelligibility. The high correlation to the word energy can be explained by irregular noise in low quality voices which are again less intelligible.

6 Conclusion and Outlook

In speech therapy and rehabilitation, a patient's voice has to be evaluated by the therapist. Speech recognition systems can be used to objectively analyze the intelligibility of pathologic speech. In this paper, the influence of reading errors and out-of-text utterances on human-machine correlation was examined. For this purpose, the effects named above were removed from the original speech samples where possible, and the correlation between speech expert and automatic speech recognizer was compared for the files with and without errors. The correlation

Table 2. Recognition results for the speech corpora with and without reading errors (85 speakers) and correlation between automatic measure and human intelligibility rating *(rightmost column)*

	measure	avg.	st. dev.	min.	max.	correl.
with errors	WA	48.0	17.2	3.4	81.3	–0.61
without errors	WA	49.3	17.0	10.1	81.3	–0.61
with errors	WR	53.2	15.3	9.1	82.2	–0.56
without errors	WR	54.1	15.4	9.1	82.2	–0.55

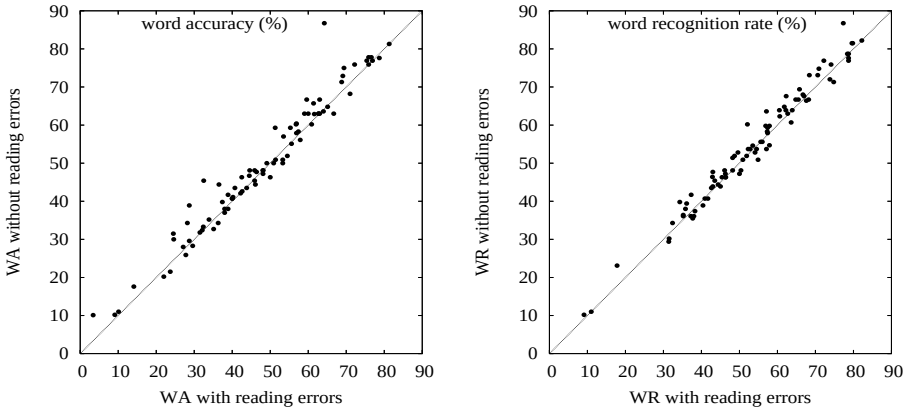

Fig. 3. Word accuracy *(left)* and word recognition rate *(right)* before and after removing reading errors; all dots above the diagonal mean better results afterward

showed no significant difference. Hence, the reading errors in text recordings do not have to be eliminated before automatic evaluation.

For the improvement of the correlation to human evaluation, adaptation of the speech recognizer to pathologic speech should be considered. However, for substitute voices after total laryngectomy, the adaptation enhances the recognition results but not the agreement to the human reference [13]. Another approach considering the recognition rates is to include the words of frequently occurring out-of-text phrases, like "I forgot my glasses.", into the recognition vocabulary of the recognizer. This is part of future work.

Acknowledgments

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

Table 3. Correlation r between selected prosodic features and human intelligibility ratings; presented are criteria with a correlation of $|r| \geq 0.5$

feature	correlation	
	with errors	without errors
ratio of duration of unvoiced segments and file length	+0.51	+0.53
duration of silent pause after current word	+0.54	+0.53
normalized energy of word-pause-word interval	+0.62	+0.59
normalized duration of word-pause-word interval	+0.63	+0.60

References

1. American Cancer Society: Cancer facts and figures 2000, Atlanta, GA (2000)
2. Makeieff, M., Barbotte, E., Giovanni, A., Guerrier, B.: Acoustic and aerodynamic measurement of speech production after supracricoid partial laryngectomy. Laryngoscope **115**(3) (2005) 546–551
3. Fröhlich, M., Michaelis, D., Strube, H.W., Kruse, E.: Acoustic voice analysis by means of the hoarseness diagram. J Speech Lang Hear Res **43**(3) (2000) 706–720
4. Schuster, M., Haderlein, T., Nöth, E., Lohscheller, J., Eysholdt, U., Rosanowski, F.: Intelligibility of laryngectomees' substitute speech: automatic speech recognition and subjective rating. Eur Arch Otorhinolaryngol **263**(2) (2006) 188–193
5. International Phonetic Association (IPA): Handbook of the International Phonetic Association. Cambridge University Press (1999)
6. Stemmer, G.: Modeling Variability in Speech Recognition. Volume 19 of Studien zur Mustererkennung. Logos Verlag, Berlin (2005)
7. Wahlster, W., ed.: Verbmobil: Foundations of Speech-to-Speech Translation. Springer, Berlin (2000)
8. 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**(5) (2000) 519–532
9. 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
10. 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
11. Batliner, A., Buckow, A., Niemann, H., Nöth, E., Warnke, V.: The Prosody Module. [7] 106–121
12. Haderlein, T.: Automatic Evaluation of Tracheoesophageal Substitute Voices. Volume 25 of Studien zur Mustererkennung. Logos Verlag, Berlin (2007)
13. Haderlein, T., Steidl, S., Nöth, E., Rosanowski, F., Schuster, M.: Automatic Recognition and Evaluation of Tracheoesophageal Speech. In Sojka, P., Kopeček, I., Pala, K., eds.: Proc. Text, Speech and Dialogue (TSD 2004). Volume 3206 of LNAI., Berlin, Springer (2004) 331–338