Can you understand him? Let's look at his word accuracy-automatic evaluation of tracheoesophageal speech

Maria Schuster, Elmar Nöth, Tino Haderlein, Stefan Steidl, Anton Batliner, Frank Rosanowski

Angaben zur Veröffentlichung / Publication details:

Schuster, Maria, Elmar Nöth, Tino Haderlein, Stefan Steidl, Anton Batliner, and Frank Rosanowski. 2005. "Can you understand him? Let's look at his word accuracy-automatic evaluation of tracheoesophageal speech." In *ICASSP'05: IEEE International Conference on Acoustics, Speech, and Signal Processing, 23-23 March 2005, Philadelphia, PA, USA*, 61–64. Piscataway, NJ: IEEE Operations Center. https://doi.org/10.1109/ICASSP.2005.1415050.

CAN YOU UNDERSTAND HIM? LET'S LOOK AT HIS WORD ACCURACY — AUTOMATIC EVALUATION OF TRACHEOESOPHAGEAL SPEECH

Maria Schuster[2], *Elmar Nöth*[1], *Tino Haderlein*[1], *Stefan Steidl*[1], *Anton Batliner*[1], *Frank Rosanowski*[2]

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

[2]Abt. für Phoniatrie und Pädaudiologie des Universitätsklinikums Erlangen
Bohlenplatz 21, 91054 Erlangen, Germany

ABSTRACT

Tracheoesophageal (TE) speech is a possibility to restore the ability to speak after laryngectomy. TE speech often shows low intelligibility. An objective means to determine and quantify the intelligibility does not exist until now and an automation of this procedure is desirable. We used a speech recognizer trained on normal, non–pathologic voices. We compared intelligibility scores for TE speech from five experienced raters with the word accuracy (WA) of our speech recognizer. A correlation coefficient of -.84 shows that WA can be a good indicator of intelligibility for pathologic voices. An outlook for future work is presented.

1. INTRODUCTION

The results of a speech recognition task depend on the quality of the input signal. The term "quality" is mostly used in this context to describe the influences of the transmission channel or background noise, but of course the speaker's voice can be the source of recognition problems as well. This paper focuses on the recognition of a special kind of pathologic voices, i.e. tracheoesophageal (TE) voices. After laryngectomy, i.e. the removal of the larynx, patients suffer from several impairments, the loss of laryngeal speech being of outstanding importance for the affected patients and their social functioning. In these patients, speech restoration can be achieved by different methods, TE techniques being increasingly popular because of their resemblance to laryngeal voice production [7]: A silicone one–way valve is placed into a shunt between the trachea and the esophagus, which on the one hand prevents aspiration and on the other hand deviates the air stream during expiration into the upper esophagus. The upper esophagus, the pharyngo–esophageal (PE) segment, serves as a sound generator (see Figure 1). Tissue vibrations of the PE segment modulate the streaming air and generate a substitute voice signal. In comparison to normal voices the quality of substitute voices is "low". Intercycle frequency perturbations result in a hoarse voice [8]. Furthermore, the change of pitch and volume is limited which causes monotone voice. Another source of distortion is the so–called tracheostoma which is the upper end of the trachea (see Figure 1). In order to force the air to take its way through the shunt into the esophagus and allow voicing, the patient usually closes the tracheostoma with a finger. If the patient is not able to do this properly, loud "whistling" noises from the eluding air occur. Acoustic studies of TE voices can be found for instance in [7, 2]. Figure 2 shows the spectrograms of the German words *"einst stritten sich"* from a TE speaker and a laryngeal speaker (the TE speaker might be considered as typical because his intelligibility was rated 2.8, which was approximately the average score across the used database, cf. below). Properties of TE speech like the low pitch and the high noise portions are clearly visible.

In this paper, we will not concentrate on acoustic properties. The reduced sound quality and problems such as the reduced ability of intonation or voiced–voiceless distinction [4, 9] lead to worse intelligibility. Although TE voices proved to be better than other substitute voices, usually patients suffer from a deteriorated quality of life, as they cannot communicate properly. Speech therapy can improve the intelligibility of a patient's TE speech. In speech therapy and rehabilitation a patient's voice has to be evaluated and measures for the description and evaluation of alaryngeal voices in laryngectomized patients are needed.

In our work we examine how well TE speech is processed by a speech recognition system and whether the results can be used for evaluating the quality of a substitute voice automatically, i.e. whether they correlate with experts' ratings.

2. THE RECOGNITION SYSTEM

The speech recognition system used for the experiments was developed at the Chair for Pattern Recognition. It can handle spontaneous speech with mid–sized vocabularies up to 10,000 words. It is used in commercial applications by *Sympalog* (www.sympalog.com), a spin–off company of our institute, for conversational speech dialogue systems. The latest version is described in detail in [3, 10].

For each frame a 24–dimensional feature vector which contains short–time energy, 11 Mel–frequency cepstral coefficients (MFCC) and their first–order derivatives is computed. The derivatives are approximated by the slope of a linear regression line over 5 consecutive frames (50 ms). The short–time analysis applies a Hamming window with a length of 16 ms, the frame rate is 10 ms. The filter bank for the Mel–spectrum consists of 25 triangle filters.

The system uses semi–continuous Hidden Markov Models (HMM). It models phones in a context as large as still statistically

This work was partly funded by the EU in the project PF–STAR under grant IST–2001–37599. The responsibility for the content lies with the authors.

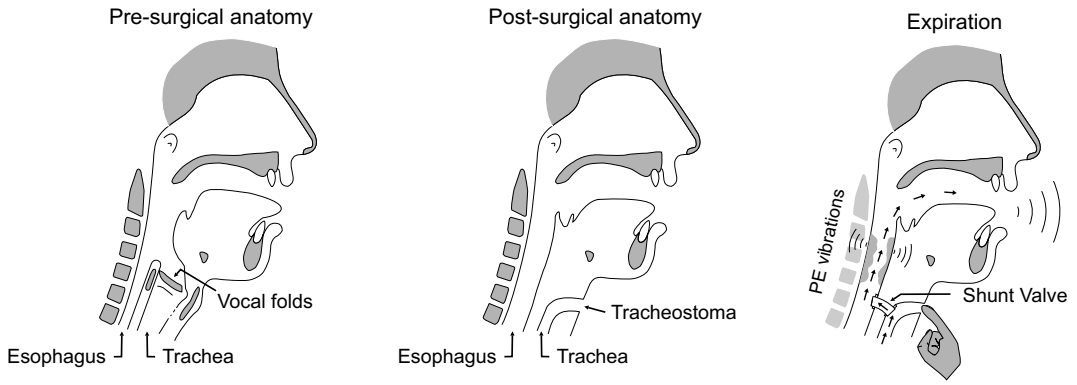

Fig. 1. Anatomy of a person with intact larynx *(left)*, anatomy after total laryngectomy *(middle)*, and the substitute voice *(right)* caused by vibration of the pharyngoesophageal segment (pictures from [6])

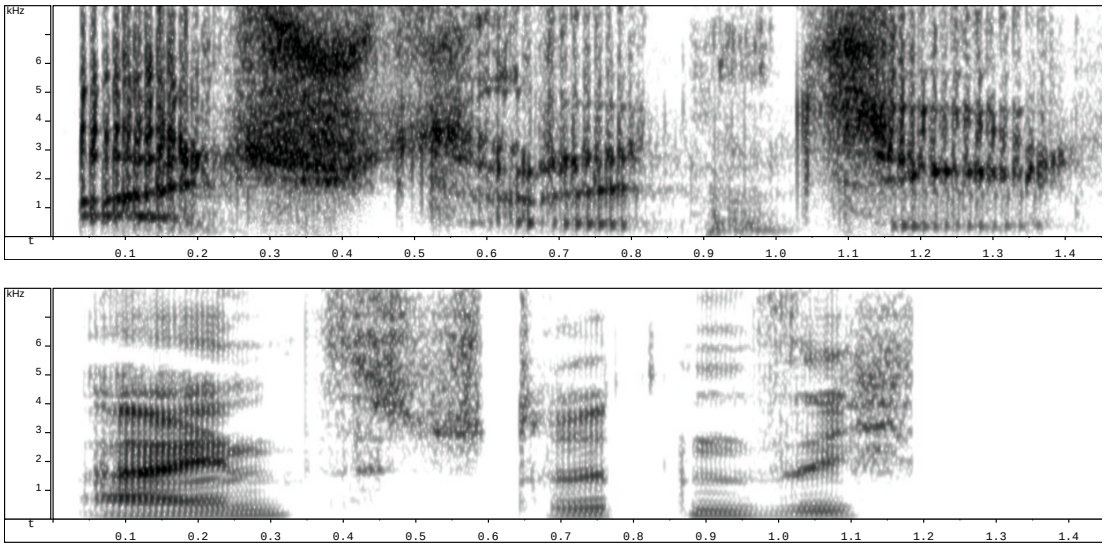

Fig. 2. Spectrograms of the German words *"einst stritten sich"* from a TE speaker *(top)* and a laryngeal speaker *(bottom)*

useful and thus forms the so–called polyphones, a generalization of the well–known bi– or triphones. The HMMs for the polyphones have three to four states. In the current experiments we use monophones, because they produce slightly better results than polyphones with the strongly distorted TE speech. The codebook has 500 classes and a unigram language model is used, so that the results are mainly dependent on the acoustic models.

3. TRAINING AND TEST DATA

The recognition system for the experiments in this paper was trained with dialogues from the VERBMOBIL project [11]. The topic in the recordings is appointment scheduling. The data were recorded with a close–talk microphone at a sampling frequency of 16 kHz and quantized with 16 bit. The speakers were from all over Germany and thus covered most dialectal regions. They were, however, asked to speak standard German. About 80% of the 578 training speakers (304 male, 274 female) were between 20 and 29 years old, less than 10% were over 40. This is important in view of the test data, because the fact that the average age of our test speakers is more than 60 years may influence the recognition results. Of the VERBMOBIL–German data (12,030 utterances, 263,633 words, 27.7 hours of speech) a subset of 11,714 utterances (257,810 words) was used for the training and 48 (1042 words) for the validation set. Thus we kept the same corpus partitions as in [3, 10].

We used three groups of speakers to test the system: 18 older male laryngectomees with TE voice were recorded ($\mu = 64.2$ years, $\sigma = 8.3$ years). They had undergone total laryngectomy because of laryngeal or hypopharyngeal cancer at least one year prior to the investigation and were provided with a Provox® shunt valve. Each person read the German version of the story "North Wind and Sun", a phonetically balanced text with 108 words (71 disjunctive) often used in German speaking countries in speech therapy. We

Table 1. Average WA for the different speaker groups

speaker group	train–matched	test–matched	TE test
mean WA	69%	58%	28%
st. dev.	10%	6%	13%

used two control groups: 18 male laryngeal speakers without laryngeal diseases or subjective voice problems (age–matched to the laryngectomees, column *test–matched* in Table 1) and 16 laryngeal speakers (age–matched to the majority of the training data, 9 male, 7 female, column *train–matched* in Table 1).

The vocabulary of the recognizer for the experiments consisted of the 71 words occurring in the test data and 32 words and syllables that served as filler models for reading errors.

Table 1 shows the word accuracy (WA) for the different speaker groups. Note that the WA is computed w.r.t. the text to be read and not w.r.t. the transliteration, i.e. we ignored reading errors (see Section 4). The results meet our expectations: The group age–matched to the training data (train-matched) is better than the group age–matched to the TE speakers (test–matched, see also [12]); the TE speakers have very bad recognition results due to the strong deviation from the training data. Note that the low recognition for the non TE speakers is due to the use of monophone models. Using polyphones instead of monophones as subword units, a WA of 84% was achieved for the train–matched group and 68% for the test–matched. However, as already mentioned, polyphones proved to be less stable in the presence of the strongly distorted TE speech.

4. HUMAN AND AUTOMATIC INTELLIGIBILITY RATING

In speech therapy and rehabilitation a patient's voice has to be evaluated by the therapist. An automatically computed, objective measure would be a very helpful support for this task. In this section we present experiments concerning the usability of WA as an objective measure.

At the Department of Phoniatrics and Paediatric Audiology at our university, five experienced voice professionals evaluated the voices of the 18 TE test persons on criteria such as "hoarseness", "prosody" and "effort", i.e. criteria that are used to characterize voice quality. The most important criterion was "intelligibility", i.e. the overall holistic evaluation of how well the patient can be understood. The scores given by the experts were represented by integers between 1 (very high) and 5 (very low). The average "intelligibility" score across all patients was 2.9 with $\sigma = 1.14$. It seemed to be obvious that a voice which is well intelligible for a human being will also achieve better results in automatic speech recognition. So we chose this single criterion and compared the experts' rating to the WA we got from our speech recognizer.

First we tested how homogeneous the expert group rated the test data. For the 18 files the correlation of each single rater's "intelligibility" scores to the average scores across the other four persons was calculated (compare Table 2). The two lowest correlation values were .68 and .77, the others were between .82 and .85. The inter–rater variance for the experts was .11. Then we measured the correlation between man and machine for the 18 recordings where the WA across a speaker's entire read story served as the automatically computed score. The results for the correlation of

Table 2. Correlation coefficients between single raters and the average of the 4 other raters for the criterion "intelligibility"

rater	K	L	R	S	U
corr.	.83	.82	.77	.85	.68

Table 3. Correlation coefficients between single raters and the average of all raters for the criterion "intelligibility" with the WA of the recognizer

rater	K	L	R	S	U	avg.
corr.	-.81	-.65	-.81	-.79	-.55	-.84

the WA to the individual expert and the average of them are shown in Table 3. Considering the average of the raters, the WA for the recognizer has a correlation of -.84. The coefficient is negative because high recognition rates came from "good" voices with a low score number and vice versa.

Figure 3 shows the WA vs. the average of the 5 experts' scores for the 18 patients with TE voice as well as the corresponding regression line; the patients are ordered w.r.t. increasing WA.

It is clearly visible that there is a strong correlation between the results of the human and the automatic analyzing method. This leads us to the assumption that the WA will be very helpful as a part of a future automatic intelligibility or, in general, voice quality analyzer.

We rounded the experts' average scores to the next integer and also mapped the WA results to the same scale. We set the thresholds on the WA results so that the difference between the experts' scores and the scores derived from WA is minimal (i.e. 0 in our case). Figure 4 shows these two scores and the applied thresholds: 12 results were identical and 6 results differed by only a grade of 1. Granted that it is unfair to optimize on the test data, the results still show that WA can be a useful measure for the analysis of TE speech.

In [5] we showed how to improve the WA with an unsupervised HMM interpolation approach from 28% to 36%. The improved WA did not lead to an improved correlation with the human raters. There we used the transliteration of the read text, because we were mainly interested in the adaptation technique and wanted to exclude the influence of reading errors. The correlation between WA and the average of the raters was -.84 in both cases, i.e. when calculating the WA w.r.t. the text to be read and w.r.t. the transliteration. Thus the results of the experiments reported here are important for the use of automatic speech recognition methods in speech therapy.

5. CONCLUSIONS AND OUTLOOK

A TE voice is a so–called substitute voice which is one possibility to give a patient back his ability to communicate by speech after laryngectomy. However, this voice which is produced in the PE segment often shows low quality and intelligibility. For 18 substitute voices an average WA of 28% was achieved. A test group of 16 laryngeal speakers who were age–matched to the majority of the training data had a WA of 69%, while a test group of 18 laryn-

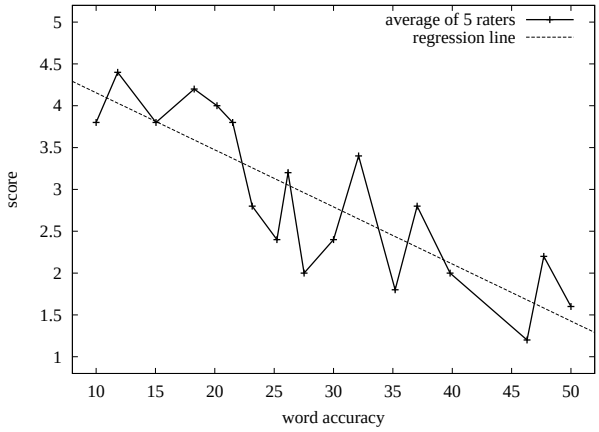

Fig. 3. WA vs. the average of the 5 experts' scores for the 18 patients with TE voice; the patients are ordered w.r.t. increasing WA.

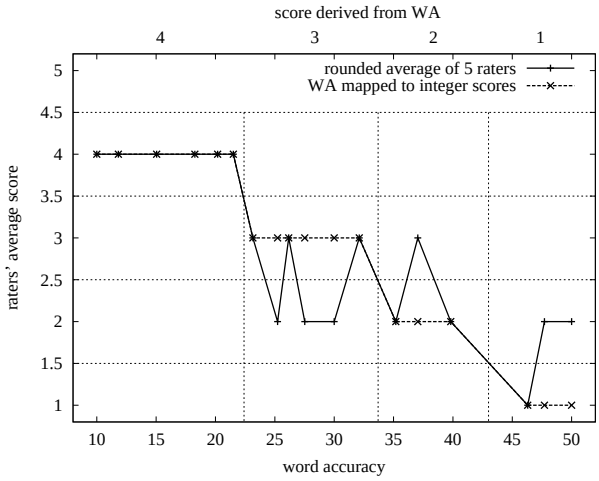

Fig. 4. Scores derived from WA vs. the rounded average of the 5 experts' scores for the 18 patients with TE voice; the patients are ordered w.r.t. increasing WA.

geal speakers who were age–matched to the TE speakers had a WA of 58%. In the field of voice evaluation we compared the intelligibility scores for recordings of TE voices from five experienced raters with the WA from our system. The monophone based recognizer's correlation was -.84 on a standard text and thus showed that an automatic evaluation of the voice quality is possible. So far we found no difference between the correlation of the experts' score and the WA w.r.t. the transliteration and w.r.t. the text to be read. However, for a future clinical application the two sources of error might have to be strictly divided. By the application of confidence measures and language models, sections with reading errors could be detected in the recording. Then the remaining parts of the file could be used for the computation of the voice quality only. More investigations have to be done with a bigger group of TE speakers to find out if such a procedure is necessary.

Besides repeating these initial experiments with a larger corpus, we plan to look at the correlation between the experts' scores and the WA for telephone TE speech. This is important because the communication via telephone is essential in modern life and it is the most difficult situation for a laryngectomee and his communication partner (absence of non–verbal communication). Also, if the correlation is good enough, the automatic evaluation could be done way more cost effective than by having to install 'analysis stations' in all clinics and rehabilitation centers.

Another aspect that we want to study is the prosody of the TE speakers. Using our prosody module [1] we want to find out, what influence the prosodic phrasing has on intelligibility. We suspect that the reduced air volume of a TE speaker forces him to pause within syntactic units which in turn has a negative effect on his intelligibility.

6. REFERENCES

[1] A. Batliner, A. Buckow, H. Niemann, E. Nöth, and V. Warnke. The Prosody Module. In Wahlster [11], pages 106–121.

[2] M.H. Bellandese, J.W. Lerman, and H.R. Gilbert. An Acoustic Analysis of Excellent Female Esophageal, Tracheoesophageal, and Laryngeal Speakers. *Journal of Speech, Language, and Hearing Research*, 44:1315–1320, 2001.

[3] F. Gallwitz. *Integrated Stochastic Models for Spontaneous Speech Recognition*, volume 6 of *Studien zur Mustererkennung*. Logos Verlag, Berlin, 2002.

[4] J. Gandour and B. Weinberg. Perception of Intonational Contrasts in Alaryngeal Speech. *Journal of Speech and Hearing Research*, 26:142–148, 1983.

[5] T. Haderlein, S. Steidl, E. Nöth, F. Rosanowski, and M. Schuster. Automatic Recognition and Evaluation of Tracheoesophageal Speech. In P. Sojka, I. Kopeček, and K. Pala, editors, *Proc. 7th International Conference on Text, Speech and Dialogue (TSD 2004)*, Lecture Notes for Artificial Intelligence, pages 331–338, Berlin, 2004. Springer–Verlag.

[6] J. Lohscheller. *Dynamics of the Laryngectomee Substitute Voice Production*. Shaker, Aachen, 2003.

[7] 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*, 49:202–210, 1984.

[8] H.K. Schutte and G.J. Nieboer. Aerodynamics of esophageal voice production with and without a Groningen voice prosthesis. *Folia Phoniatrica et Logopaedia*, 54:8–18, 2002.

[9] J.P. Searl and M.A. Carpenter. Acoustic Cues to the Voicing Feature in Tracheoesophageal Speech. *Journal of Speech, Language, and Hearing Research*, 45:282–294, 2002.

[10] G. Stemmer. *Modeling Variability in Speech Recognition*. PhD thesis, Chair for Pattern Recognition, University of Erlangen-Nuremberg, 2004.

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

[12] J.G. Wilpon and C.N. Jacobsen. A study of speech recognition for children and the elderly. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 349–352, 1996.