An Automatic Version of the Post-Laryngectomy Telephone Test

Tino Haderlein[1,2], Korbinian Riedhammer[1], Andreas Maier[1,2], Elmar Nöth[1], Hikmet Toy[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. Tracheoesophageal (TE) speech is a possibility to restore the ability to speak after total laryngectomy, i.e. the removal of the larynx. The quality of the substitute voice has to be evaluated during therapy. For the intelligibility evaluation of German speakers over telephone, the Post-Laryngectomy Telephone Test (PLTT) was defined. Each patient reads out 20 of 400 different monosyllabic words and 5 out of 100 sentences. A human listener writes down the words and sentences understood and computes an overall score. This paper presents a means of objective and automatic evaluation that can replace the subjective method. The scores of 11 naïve raters for a set of 31 test speakers were compared to the word recognition rate of speech recognizers. Correlation values of about 0.9 were reached.

1 Introduction

In 20 to 40 percent of all cases of laryngeal cancer, total laryngectomy has to be performed, i.e. the removal of the entire larynx [1]. For the patient, this means the loss of the natural voice and thus the loss of the main means of communication. One possibility of voice restoration is the tracheoesophageal (TE) substitute voice. In TE speech, the upper esophagus, the pharyngo-esophageal (PE) segment, serves as a sound generator (see Fig. 1). The air stream from the lungs is deviated into the esophagus during expiration via a shunt between the trachea and the esophagus. Tissue vibrations of the PE segment modulate the streaming air and generate a substitute voice signal. 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. In comparison to normal voices, the quality of substitute voices is "low". Inter-cycle frequency perturbations result in a hoarse voice [2]. Furthermore, the change of pitch and volume is limited which causes monotone voice. Acoustic studies of TE voices can be found for instance in [3, 4]. The reduced sound quality and problems such as the reduced

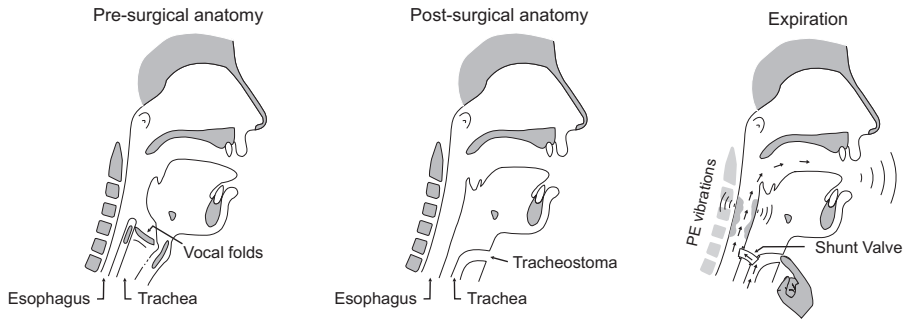

Fig. 1. Anatomy of a person with intact larynx *(left)*, anatomy after total laryngectomy *(middle)*, and the substitute voice *(right)* caused by vibration of the pharyngo-esophageal segment (pictures from [10])

ability of intonation or voiced-voiceless distinction [5, 6] lead to worse intelligibility. For the patients, this means a deterioration of quality of life as they cannot communicate properly.

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 current automatic evaluation, usually sustained vowels are analyzed and the voice quality is rated. However, for criteria like intelligibility not just a voice sample but a speech sample is needed. Moerman et al. [7] investigated recordings of a short text that contained 18 words. Correlations to human ratings were only given for the "overall impression" of the substitute voice ($r = 0.49$), so no direct comparisons to our study are possible. In previous work, we showed that an automatic speech recognition (ASR) system can be used to rate the intelligibility in close-talking speech of post-laryngectomy speakers [8, 9]. The telephone is a crucial part of the patients' social life, and it is necessary for them to have a means of communication that does not require them to leave their home. Therefore, intelligibility on a telephone reflects an everyday communication situation which is important for the patient. In this paper, we will present an automatic version of an introduced standard test for intelligibility over the telephone.

This paper is organized as follows: In Sect. 2, the Post-Laryngectomy Telephone Test will be explained. The test data will be introduced in Sect. 3. Section 4 will give some information about the speech recognition system. Section 5 contains the results, and Sect. 6 will give a short outlook on future work.

2 The Post-Laryngectomy Telephone Test

The Post-Laryngectomy Telephone Test (PLTT, [11]) was developed in order to represent the communication situation outside the patient's usual environment (the family) by taking into account both voice and language. The patient

calls a naïve rater over a standard landline telephone. The rater should not know about the text material of the test and may not have any hearing impairment.

The PLTT vocabulary consists of 400 monosyllabic words and 100 sentences, each of them written on an individual card. For one session, 22 words and 6 sentences are randomly chosen. The first two words and the first sentence are not taken into account for evaluation. Instead, they are supposed to allow the listener to adapt to the speaker. The speaker may only read what is written on the cards. Any further utterances, like e.g. articles (the German language has different ones for each grammatical gender), are not allowed. The test begins with reading the words. When the listener does not understand a word, he or she may say exactly once: "Please repeat the word." Further feedback about the intelligibility is not allowed. The sentences may not be repeated.

Three measures are computed from the listening experiment. The number of words w the listener understood correctly during the first attempt is multiplied by 5 and represents the word intelligibility i_{word} in percent. Words that were repeated do not get a point. Each sentence s gets a score c_s of 0 to 2 points. Two points are assigned when the sentence was understood completely correct. One point is given if one word is missing or not understood correctly. In all other cases, the reader gets no point. The sentence intelligibility i_{sent} in percent is the resulting sum of points multiplied with 10. The total intelligibility i_{total} is then given by

$$i_{\text{total}} = \frac{i_{\text{word}} + i_{\text{sent}}}{2} = \frac{1}{2}\left(5w + 10\sum_{s=1}^{5} c_s\right) \quad . \tag{1}$$

The test was shown to be valid, reliable and objective [11], and it was also applied to laryngectomized persons before: Patients with shunt valves reached an average PLTT result between 70 and 80 [12]. A reason why the test should be done via telephone was also given: A quiet room does not reflect a real-world communication situation as noise is present almost everywhere. In a noise-free environment, the voice rehabilitation progress would be overestimated. The telephone situation is easy to maintain and thus suitable for practical use. But like each evaluation that involves human raters, this test is subjective for many reasons, like the listener's hearing abilities or experience with TE voices. Other persons might not be able to understand or reproduce the results. For these reasons, an objective and automatic version of the PLTT using automatic speech recognition was desired.

3 Test Data

A test set of PLTT recordings (*pltt_8kHz*) from 31 laryngectomees was available where each recording contained all words and sentences the respective speaker read out. The speakers were 25 men and 6 women (63.4±8.7 years old) with tracheoesophageal substitute speech. They were provided with a shunt valve of the Provox® type [13]. The data were recorded with a dialogue system provided by Sympalog Voice Solutions[3]. The audio files were also segmented by hand so

[3] http://www.sympalog.com

that each word and sentence was stored in a separate file. This was done in order to explore whether the automatic evaluation is influenced by noise or non-verbals between the words in the full recordings. This database is denoted as *pltt_seg_8kHz*.

The human listeners were 8 male and 3 female students (average age: 22.5±1.2 years). None of them had experience with voice and speech analysis. For recording the PLTT, each patient got a unique sheet of paper with instructions and 22 words and 6 sentences that were randomly chosen. The first two words and the first sentence were neither used for human nor for automatic evaluation. The raters listened to the *pltt_seg_8kHz* data set. They could pause the play-back to write down the understood utterance.

4 The Speech Recognition System

The speech recognition system used for the experiments was developed at the Chair of Pattern Recognition in Erlangen. It can handle spontaneous speech with mid-sized vocabularies up to 10,000 words. The latest version is described in detail in [14]. 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; for the PLTT experiments, 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 zerogram language model was used so that the results are only dependent on the acoustic models.

The baseline system for the experiments in this paper was trained with German dialogues from the VERBMOBIL project [15]. 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. The speakers were from all over Germany and covered most regions of dialect. 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. 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 [14].

A speech recognition system can only recognize the words stored in its vocabulary list. This list had to be created from the words and sentences occurring in the PLTT. This, however, is not enough to simulate the human listener. A hu-

man being knows more words than those occurring in the test which might cause misperceptions. In order to simulate this in the automatic test, the vocabulary list of the recognizer had to be extended by words phonetically similar to those of the actual vocabulary. This was done by the definition of a modified Levenshtein distance for phonetic transcriptions. It involved a weighting function which assigns phoneme pairs that sound similar (e.g. /s/ and /z/) a low weight and thus finds the desired words [16]. In this way, the basic PLTT vocabulary that consisted of 738 words (*PLTT-small*) was expanded to 1017 words (*PLTT-large*). The additional words and their transliterations were taken from the CELEX dictionary [17]. The VERBMOBIL baseline training set was downsampled to 8 kHz sampling frequency, a VERBMOBIL recognizer was trained and the vocabulary changed to the *PLTT-small* or *PLTT-large* word list, respectively. For both cases, a polyphone-based and a monophone-based recognizer version were created. Monophones were supposed to be more robust for recognition of highly pathologic speech because each of them is trained with more data than a polyphone model.

5 Results

Table 1 shows the PLTT results of the single raters. Although they had never heard TE voices before, the inter-rater correlation for the total intelligibility i_{total} is greater than 0.8 for all persons. However, perceptive results vary strongly among the raters. The difference in the average of i_{total} for the "best" and the "worst" rater is more than 20 points which shows how strongly the test depends on the particular listener. The standard deviation is very similar for all raters, however. The recognition results and the PLTT measures both for recognizers and human raters are assembled in Table 2. Since the first part of a PLTT session consists of single words, not only the word accuracy (WA) but also the word recognition rate (WR) was computed. It is based on the word accuracy, but the number of words wrongly inserted by the recognizer is not counted. In comparison to the human WA which reached 55%, the automatic recognition rates are much lower due to the following reasons: First of all, the recognizers were trained with normal speech. This simulates a naïve listener who has not heard TE voices before, i.e. the kind of listener that is required for the PLTT. The average WA for close-talking recordings of laryngectomees was determined at approx. 30% [8, 9]. Here, the results are even lower: The speakers had read another text right before the PLTT and were therefore exhausted. The bad signal quality of the telephone transmission and the fact that the training data of the recognizers were just downsampled and not real telephone speech had also negative influence. No sentence was recognized completely correct according to the PLTT rules. For this reason, i_{sent} was 0 for all recognizers. WA and WR for the human raters were computed from the raters' written transliteration of the audio files.

Although the automatic recognition yielded so bad results, the correlation to the human ratings was high (see Table 3). The reason is that the crucial measure is not the average of the recognition rate but its range or standard de-

Table 1. Human evaluation results i_{word}, i_{sent} and i_{total}; Pearson's correlation r and Spearman's correlation ρ are calculated between the respective rater and the average of the remaining 10 raters

	i_{word}				i_{sent}				i_{total}			
rater	μ	σ	r	ρ	μ	σ	r	ρ	μ	σ	r	ρ
BM	31.4	21.5	0.91	0.90	48.1	30.4	0.88	0.88	39.8	22.6	0.92	0.90
BT	29.2	20.7	0.83	0.80	52.2	29.5	0.88	0.87	40.7	21.8	0.90	0.90
CV	43.9	20.1	0.90	0.90	45.9	28.6	0.83	0.81	44.9	22.5	0.89	0.89
GM	39.4	21.4	0.87	0.83	48.1	28.9	0.88	0.86	43.8	23.3	0.93	0.93
HT	43.0	21.3	0.89	0.84	57.2	28.6	0.77	0.76	50.1	22.4	0.86	0.86
KC	41.7	20.8	0.93	0.91	46.9	28.8	0.65	0.60	44.3	20.8	0.85	0.83
MM	47.2	23.5	0.91	0.90	50.0	28.7	0.79	0.78	48.6	23.9	0.92	0.92
PC	34.1	21.9	0.87	0.83	48.4	29.5	0.87	0.85	41.3	22.4	0.92	0.92
SM	51.1	21.3	0.82	0.82	59.4	24.0	0.82	0.75	55.2	19.2	0.90	0.86
ST	53.6	23.0	0.92	0.92	67.5	29.6	0.72	0.64	60.6	23.4	0.86	0.85
WW	40.3	19.2	0.89	0.90	56.6	25.2	0.87	0.89	48.4	19.7	0.94	0.94

viation, respectively. It was not the task of the experiments to optimize the mean recognition rate. For this reason, voices with low quality often receive negative values of WA. Nevertheless, the distribution of these values corresponds well to the measures obtained by the human listeners. The best correlation between an automatic measure and the overall PLTT result i_{total} was reached for WR on the polyphone-based recognizers. Both Pearson's correlation r and Spearman's correlation ρ are about 0.9.

The outcome of these experiments is that the PLTT can be replaced by an objective, automatic approach. The question whether monophone-based or polyphone-based recognizers are better for the task could not be answered. When the word accuracy WA was compared to i_{total}, monophones were advantageous; when the word recognition rate WR was used instead, the polyphone-based recognizers were closer to the human rating. There are also some cases in which the correlation is slightly better when each word and sentence is processed separately, but in general the long *pltt_8kHz* recordings which contain the entire test can be used without prior segmentation.

6 Conclusion and Outlook

In this paper, an approach for the automation of the Post-Laryngectomy Telephone Test (PLTT) was presented. The correlation between the overall intelligibility score that is usually computed by a human listener and the word recognition rate of a speech recognizer was about 0.9. A difference between the human and the machine evaluation was that the automatic version does not process words again it did not "understand" on first attempt. This is not necessary since the result would be the same. Furthermore, a word that was not understood by the listener on first attempt does not get a point anyway, so it is not necessary to consider word repetition in the automatic version at all.

Table 2. Average word accuracy (WA), word recognition rate (WR), and the PLTT measures i_{word}, i_{sent} and i_{total} for speech recognizers and human raters

data set	*pltt_8kHz*				*pltt_seg_8kHz*				
vocabulary	*PLTT-small*		*PLTT-large*		*PLTT-small*		*PLTT-large*		raters
recog. units	*mono*	*poly*	*mono*	*poly*	*mono*	*poly*	*mono*	*poly*	
μ(WA)	10.0	1.8	8.0	–0.1	9.2	0.3	7.4	–1.5	55.1
σ(WA)	14.8	20.4	13.5	19.9	14.7	21.4	12.9	20.2	21.4
μ(WR)	17.3	16.6	14.4	13.7	16.4	15.6	14.2	13.2	55.3
σ(WR)	13.2	12.6	9.3	11.2	9.9	10.8	8.7	10.3	21.4
$\mu(i_{\text{word}})$	17.8	13.1	14.5	10.9	14.1	11.1	12.0	9.4	41.4
$\sigma(i_{\text{word}})$	15.1	13.0	12.8	10.9	13.8	11.6	12.7	11.1	21.3
$\mu(i_{\text{sent}})$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	52.8
$\sigma(i_{\text{sent}})$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	28.3
$\mu(i_{\text{total}})$	8.9	6.6	7.3	5.5	7.0	5.5	6.0	4.7	47.1
$\sigma(i_{\text{total}})$	7.5	6.5	6.4	5.8	6.9	5.8	6.4	5.6	22.0

Adaptation of the speech recognizers to the signal quality might enhance the recognition results. Experiments in order to find out whether also the correlation to the human results will get better will be part of future work. Another aspect that will be taken into consideration are reading errors by the patient that have to be identified before the intelligibility measure is computed.

Acknowledgments

This work was partially funded by the German Cancer Aid (Deutsche Krebshilfe) under grant 106266. The responsibility for the contents of this study lies with the authors.

References

1. van der Torn, M., Mahieu, H., Festen, J.: Aero-acoustics of silicone rubber lip reeds for alternative voice production in laryngectomees. J Acoust Soc Am **110**(5 Pt 1) (2001) 2548–2559
2. Schutte, H., Nieboer, G.: Aerodynamics of esophageal voice production with and without a Groningen voice prosthesis. Folia Phoniatr Logop **54**(1) (2002) 8–18
3. Robbins, J., Fisher, H., Blom, E., Singer, M.: A Comparative Acoustic Study of Normal, Esophageal, and Tracheoesophageal Speech Production. J Speech Hear Disord **49**(2) (1984) 202–210
4. Bellandese, M., Lerman, J., Gilbert, H.: An Acoustic Analysis of Excellent Female Esophageal, Tracheoesophageal, and Laryngeal Speakers. J Speech Lang Hear Res **44**(6) (2001) 1315–1320
5. Gandour, J., Weinberg, B.: Perception of Intonational Contrasts in Alaryngeal Speech. J Speech Hear Res **26**(1) (1983) 142–148
6. Searl, J., Carpenter, M.: Acoustic Cues to the Voicing Feature in Tracheoesophageal Speech. J Speech Lang Hear Res **45**(2) (2002) 282–294

Table 3. Pearson's correlation r and Spearman's correlation ρ between the speech recognizers' results ("rec") and the human raters' average values ("hum")

data set	*pltt_8kHz*				*pltt_seg_8kHz*			
vocabulary	*PLTT-small*		*PLTT-large*		*PLTT-small*		*PLTT-large*	
recognition units	*mono*	*poly*	*mono*	*poly*	*mono*	*poly*	*mono*	*poly*
$r(\mathrm{WA_{rec}}, \mathrm{WA_{hum}})$	0.72	0.71	0.73	0.70	0.73	0.67	0.71	0.69
$\rho(\mathrm{WA_{rec}}, \mathrm{WA_{hum}})$	0.85	0.82	0.83	0.82	0.83	0.79	0.80	0.80
$r(\mathrm{WR_{rec}}, \mathrm{WA_{hum}})$	0.82	0.86	0.81	0.83	0.87	0.88	0.86	0.87
$\rho(\mathrm{WR_{rec}}, \mathrm{WA_{hum}})$	0.88	0.94	0.89	0.92	0.91	0.91	0.89	0.91
$r(\mathrm{WA_{rec}}, i_{\mathrm{total,hum}})$	0.71	0.72	0.72	0.71	0.72	0.67	0.71	0.70
$\rho(\mathrm{WA_{rec}}, i_{\mathrm{total,hum}})$	0.84	0.81	0.83	0.80	0.81	0.76	0.79	0.79
$r(\mathrm{WR_{rec}}, i_{\mathrm{total,hum}})$	0.81	0.88	0.82	0.85	0.85	0.89	0.86	0.89
$\rho(\mathrm{WR_{rec}}, i_{\mathrm{total,hum}})$	0.86	0.93	0.87	0.92	0.88	0.90	0.90	0.90

7. Moerman, M., Pieters, G., Martens, J., van der Borgt, M., Dejonckere, P.: Objective evaluation of the quality of substitution voices. Eur Arch Otorhinolaryngol **261**(10) (2004) 541–547
8. Schuster, M., Nöth, E., Haderlein, T., Steidl, S., Batliner, A., Rosanowski, F.: Can You Understand Him? Let's Look at His Word Accuracy – Automatic Evaluation of Tracheoesophageal Speech. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). Volume I., Philadelphia, PA (USA) (2005) 61–64
9. 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
10. Lohscheller, J.: Dynamics of the Laryngectomee Substitute Voice Production. Volume 14 of Berichte aus Phoniatrie und Pädaudiologie. Shaker Verlag, Aachen (Germany) (2003)
11. Zenner, H.: The postlaryngectomy telephone intelligibility test (PLTT). In Herrmann, I., ed.: Speech Restoration via Voice Prosthesis. Springer, Berlin, Heidelberg (1986) 148–152
12. de Maddalena, H., Zenner, H.: Evaluation of speech intelligibility after prosthetic voice restoration by a standardized telephone test. In Algaba, J., ed.: Proc. 6th International Congress on Surgical and Prosthetic Voice Restoration After Total Laryngectomy, San Sebastian (Spain), Elsevier Science (1996) 183–187
13. Hilgers, F., Balm, A.: Long-term results of vocal rehabilitation after total laryngectomy with the low-resistance, indwelling Provox voice prosthesis system. Clin Otolaryngol **18**(6) (1993) 517–523
14. Stemmer, G.: Modeling Variability in Speech Recognition. Volume 19 of Studien zur Mustererkennung. Logos Verlag, Berlin (Germany) (2005)
15. Wahlster, W., ed.: Verbmobil: Foundations of Speech-to-Speech Translation. Springer, Berlin (Germany) (2000)
16. Riedhammer, K.: An Automatic Intelligibility Test Based on the Post-Laryngectomy Telephone Test. Student's thesis, Lehrstuhl für Mustererkennung (Chair for Pattern Recognition), Universität Erlangen–Nürnberg, Erlangen (Germany) (2007)
17. R.H. Baayen, R. Piepenbrock, and L. Gulikers. *The CELEX Lexical Database (Release 2).* Linguistic Data Consortium, Philadelphia, PA (USA), 1996.