Automatic Evaluation of Tracheoesophageal Telephone Speech

Korbinian Riedhammer[†], Tino Haderlein[*], Maria Schuster[*], Frank Rosanowski[*], Elmar Nöth[†]

[*] Abteilung für Phoniatrie und Pädaudiologie, Universität Erlangen–Nürnberg
Bohlenplatz 21, 91054 Erlangen, Germany

[†] Lehrstuhl für Mustererkennung, Universität Erlangen–Nürnberg
Martensstraße 3, 91058 Erlangen, Germany
noeth@informatik.uni-erlangen.de

Abstract

The tracheoesophageal (TE) substitute voice is currently state–of–the–art treatment to restore the ability to speak after laryngectomy. The intelligibility while talking over a telephone is an important clinical factor, as it is a crucial part of the patients' social life. An objective way to rate the intelligibility of substitute voices when talking over a telephone is desirable to improve the post–laryngectomy speech therapy. An automatic speech recognition (ASR) system was applied to 41 high quality recordings of post–laryngectomy patients. The ASR system was trained with normal, non–pathologic speech. It yielded a word accuracy (WA) of 36.9%±18.0%; compared to the intelligibility rating of a group of human experts the ASR system had a correlation coefficient of -.88. After downsampling the 41 recordings to telephone quality, the ASR system reached a WA of 26.4%±13.9% leading to a correlation coefficient of -.80. These results confirm that an ASR system can be used for objective intelligibility rating over the telephone.

Samodejna evalvacija traheoezofagalnega telefonskega govora

Traheoezofagalni nadomestni glas je trenutno najsodobnejši način obnove sposobnosti govora po laringektomiji. Razumljivost pri telefonskem pogovoru je pomemben kliničen dejavnik, saj predstavlja ključen del pacientove socialne interakcije. Za izboljšanje govorne terapije po laringektomiji je zaželen objektiven način ocenjevanja razumljivosti nadomestnih glasov pri telefonskem pogovoru. S sistemom za samodejno razpoznavanje govora (SRG) je bilo pregledanih 41 visoko kakovostnih posnetkov pacientov po laringektomiji. Sistem SRG so učili z normalnim, nepatološkim govorom. Odstotek pravilno razpoznanih besed je bil 36,9%±18,0%; v primerjavi z ocenami razumljivosti, ki jih je podala skupina strokovnjakov, je imel sistem SRG korelacijski koeficient -,88. Po znižanju frekvence vzorčenja 41 posnetkov na telefonsko kakovost je sistem SRG dosegel naslednji odstotek pravilno razpoznanih besed: 26,4%±13,9% oziroma korelacijski koeficient -,80. Ti rezultati potrjujejo, da je sistem SRG primeren za objektivno ocenjevanje razumljivosti telefonskega govora.

1. Introduction

The tracheoesophageal (TE) substitute voice is currently state–of–the–art treatment to restore the ability to speak after laryngectomy (Brown et al., 2003): 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 the primary substitute voice signal which is then further modulated in the same way as normal speech. In comparison to normal voices the quality of substitute voices is low, e.g. the change of pitch and volume is limited and inter–cycle frequency perturbations result in a hoarse voice (Schutte and Nieboer, 2002). Another source of distortion is the so–called tracheostoma which is at 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 (Robbins et al., 1984; Bellandese et al., 2001).

In order to improve post–laryngectomy speech therapy, an objective means to rate intelligibility is desired. In previous work we showed that an automatic speech recognition (ASR) system can be used to rate the intelligibility (Schuster et al., 2006; Schuster et al., 2005) of post–laryngectomy speakers. As the telephone is a crucial part of the patients' social life, an objective rating of the intelligibility when talking over a telephone would enhance post–laryngectomy speech therapy.

In our work we examine how well TE telephone speech is processed by an ASR system and how we can optimize the recognition system to achieve better results in order to provide a proper objective intelligibility measure for telephone data.

2. The Recognition System

The ASR system used for the experiments was developed at the Chair of Pattern Recognition (Lehrstuhl für Mustererkennung) of the University of Erlangen–Nuremberg. It can handle spontaneous speech with mid–sized vocabularies up to 10,000 words. A commercial version of this recognizer is used in high–end telephone–based conversational dialogue systems by *Sympalog* (www.sympalog.com), a spin–off company of the Chair of Pattern Recognition. The latest version is described in detail in (Gallwitz, 2002; Stemmer, 2005).

The short–time analysis applies a Hamming window with a length of 16 ms, the frame rate is 10 ms. For each frame, a 24–dimensional feature vector is computed which

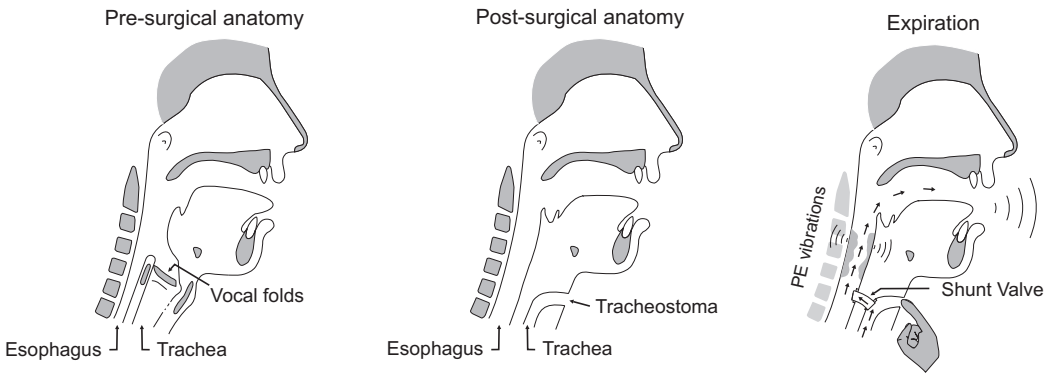

Figure 1: Physiological changes and speaking after laryngectomy: Anatomy of a person with intact larynx (*left*), anatomy after total laryngectomy (*middle*), and the substitute voice (*right*) caused by vibration of the pharyngoesophageal segment (pictures from (Lohscheller, 2003)).

contains the short–time energy, 11 Mel–frequency cepstral coefficients (MFCC) and their first–order derivatives. The derivatives are approximated by the slope of a linear regression line over 5 consecutive frames (56 ms). The filter bank for the Mel–spectrum consists of 25 triangle filters. The actual recognition is done using semi–continuous Hidden Markov Models (SCHMMs). The codebook contains 500 Gaussian densities which are shared by all HMM states. Also, a unigram language model is used, so that the results are mainly dependent on the acoustic models. The elementary recognition units are polyphones, an extension of the well–known triphone approach (Schukat-Talamazzini, 1995). The HMMs for the polyphones have three to four states.

3. Recognizer Training

The basic training set for our recognizers are dialogues from the VERBMOBIL project (Wahlster, 2000). The topic of 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 (linear). The speakers were from all over Germany and thus covered most dialectical regions. However, they were 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. A subset of the German VERBMOBIL data (11,714 utterances, 257,810 words, 25 hours of speech) was used for the training set and 48 utterances (1042 words) for the validation set[1].

In order to get a telephone speech recognizer, we downsampled the training set to telephone quality. We reduced the sampling rate to 8 kHz and applied a low–pass filter with a cutoff frequency of 3400 Hz to simulate telephone quality.

In (Schuster et al., 2005), we showed for a corpus of 18 TE speakers that a monophone–based recognizer for close–talk signals produced slightly better agreement with speech experts' intelligibility ratings than a polyphone–based recognizer. We wanted to verify these results for a larger corpus. Therefore we created four different recognizers: For the 16 kHz and the 8 kHz training data, we created a polyphone–based and a monophone–based recognizer (rows "16kHz/mono", "8kHz/mono", "16kHz/poly", "8kHz/poly" in Table 3). After the training, the vocabulary was reduced to the words occurring in the German version of the "The North Wind and the Sun" text, a fable from Aesop. It is a phonetically rich text with 108 words (71 disjoint) which is often used in speech therapy in German speaking countries.

4. Evaluation Data

41 laryngectomees ($\mu = 62.0 \pm 7.7$ years old, 2 female and 39 male) with TE substitute voice read the German version of the text "The North Wind and the Sun". The speech samples were recorded with a close–talk microphone ("dnt Call 4U Comfort" headset) at a sampling frequency of 16 kHz and quantized with 16 bit (linear).

Eight of the patients additionally read the "The North Wind and the Sun" text to an automatic telephone–based recording system (the recording system was not yet available at the time of the recording of the other 33 patients). The samples were recorded with 8 kHz and quantized with 16 bit (linear). However, one has to keep in mind that the signal is logarithmically companded (8 bit) during transmission which is approximately equivalent to 12 bit linear (rows "telephone calls" in Table 3).

Each close–talk recording was rated by 5 voice professionals (see Sec. 5.). Previous work (Schuster et al., 2006; Schuster et al., 2005) showed that there exists a significant correlation between experts' intelligibility ratings and the speech recognizer's word accuracy (WA) for close–talk recordings. If an automatic evaluation of TE telephone speech is possible, there must be a similar correlation using telephone data. To determine the change of correlation, we created three additional versions of the close–talk data:

1. We downsampled the data to 8 kHz applying the same low–pass filter (3400 Hz) as for the training data (rows

[1]The training and validation corpus was thus the same as in (Gallwitz, 2002; Stemmer, 2005).

"low–pass 3400" in Table 3).

2. In order to simulate the loss due to the logarithmic encoding in the telephone channel, we converted these linearly quantized signals to μ–law companded signals and back to linearly quantized signals (rows "low–pass 3400, μ–law" in Table 3).

3. In order to get a "telephone quality" version of the signals, we played back the close–talk recordings using a standard PC and loudspeaker in a quiet office environment and placed a telephone headset in front of the loudspeaker. The replayed sound files were recorded with the same automatic dialogue system over the telephone mentioned above with 8 kHz and 16 bit linear (again, the signals were logarithmically companded during telephone transmission). Thus we simulated a real telephone call (rows "simulated telephone" in Table 3). Due to the multiple AD/DA conversions and the different frequency characteristics of the loudspeaker and the microphones we expect the recognition rates to be a lower bound for the recognition rates for real telephone calls.

Figure 2 shows spectrograms of a short passage from the "The North Wind and the Sun" fable. The recordings are from one speaker who was recorded with the close–talk microphone (top) and with the telephone–based system (bottom). The spectrogram in the middle is from the downsampled close–talk version which was μ–law companded.

5. Subjective Evaluation

A group of 5 voice professionals subjectively estimated the intelligibility of the patients while listening to a play–back of the close–talk recordings. A five–point Likert scale (1 = very high, 2 = rather high, 3 = medium, 4 = rather low, 5 = very low) was applied to rate the intelligibility of each recording. In this manner an averaged mark – expressed as a floating point value – for each patient could be calculated.

To judge the agreement between the different raters we calculated correlation coefficients and the weighted multi–rater κ. For each rater we calculated the correlation between his "intelligibility" rating and the average of the 4 other raters. Table 1 shows the correlation coefficient for each rater and the average correlation coefficient.

rater	K	L	R	S	U	avg.
others	.82	.80	.81	.85	.77	.81

Table 1: Correlation coefficients between single raters and the average of the 4 other raters for the criterion "intelligibility".

The weighted multi–rater κ by Davies and Fleiss (Davies and Fleiss, 1982) also allows to compare an arbitrary number of raters and weights the difference between the values to compare. This means e.g. for the case that rater a gives a score of 2 and rater b gives a score of 3, this pair of numbers "matches better" and is therefore weighted higher as if person b rated the test data with a 4.

The weights were chosen as proposed by Cicchetti (Cicchetti, 1976) with

$$w_{xy}^{(a,b)} = 1 - \left(\frac{x-y}{C-1}\right)^2 \quad . \tag{1}$$

A κ value greater than $.4$ is said to show moderate agreement. The weighted multi–rater κ for the 5 raters was .45.

6. Automatic Evaluation

We used the experts' intelligibility ratings for the close–talk recordings as a reference for all 4 versions of the recordings: We applied the two close–talk recognizers and the two telephone speech recognizers to the accordant speech data and calculated the correlation between the WAs and the average of the experts' intelligibility rating. The κ values were calculated using the recognizer as a 6th rater. For this we mapped the WAs to marks on the Likert scale, using the thresholds that are given in Table 2.

WA	< 0	< 15	< 25	< 40	≥ 40
Mark	5	4	3	2	1

Table 2: Thresholds for mapping the WA of the ASR system to marks on the Likert scale for rating the intelligibility of the patients.

Table 3 shows the results for the monophone–based recognizers (row 1–4) and the polyphone–based recognizers (row 5–8) for the 41 patients. In addition, the results for the 8 real telephone calls are displayed (row 9-10). Note that the correlation and κ value were computed w.r.t. the ratings of the close–talk data of these patients, i.e. a different recording. The WA for these 8 patients was 23.0% for the simulated telephone calls and and 39.7% for the close–talk recordings using the polyphone–based recognizer compared to 37.0 for the real telephone calls.

Figure 3 shows the WAs of the 41 close–talk recordings compared to the simulated telephone recordings using polyphone–based recognizers. The recordings are ordered with increasing WA for the close–talk recordings.

Figure 4 shows for the 41 recordings the WA in comparison to the average of the experts' intelligibility scores using simulated telephone data and the polyphone–based recognizer.

7. Discussion

The results of the evaluation for the 41 patients show the possibility of an automatic objective way to rate the intelligibility of TE speech. The correlation between the WA of the respective polyphone–based recognizers and the average of the experts' intelligibility scores is only reduced from -.88 to -.80, when going from close–talk to simulated telephone speech.

Adding the recognizer as a 6^{th} expert to the expert group, does not change the κ value significantly. Due to the loss of quality in telephone transmission, the multiple AD/DA conversions, and the different frequency characteristics of the loudspeaker and the microphones, the overall WA for the simulated telephone calls is reduced. Also, the

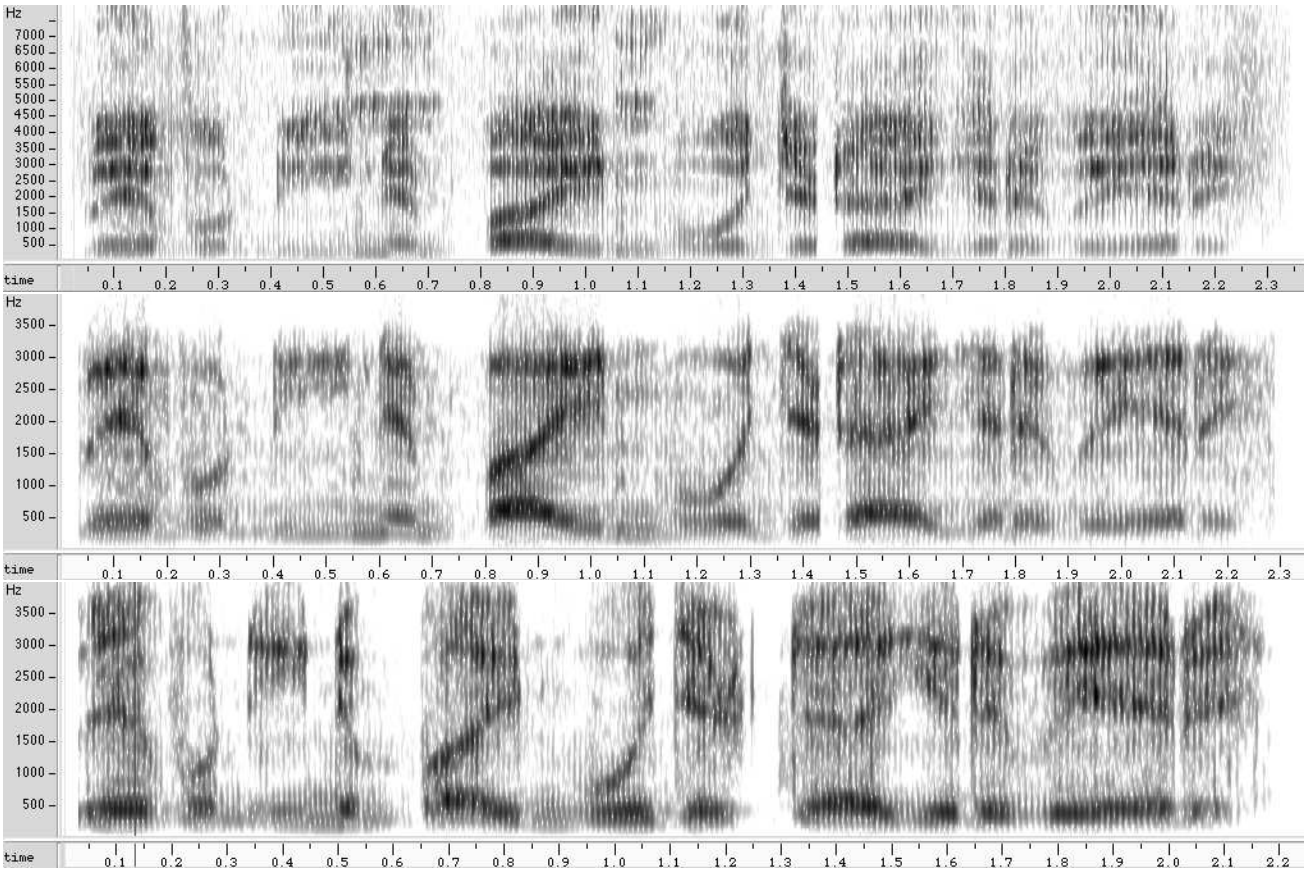

Figure 2: Spectrograms from the German utterance “wer von ihnen beiden wohl der Stärkere wäre”: 16 kHz close–talk vs. 8 kHz downsampled and μ–law companded vs. 8 kHz real telephone data.

#	recording	data/recognizer	μ(WA)	σ(WA)	correlation	weighted κ
41	close–talk	16kHz/mono	35.3	13.7	-.82	.41
41	low–pass 3400	8kHz/mono	33.4	12.1	-.81	.42
41	low–pass 3400, μ–law	8kHz/mono	33.6	12.7	-.78	.42
41	simulated telephone	8kHz/mono	28.4	10.3	-.69	.42
41	close–talk	16kHz/poly	36.9	18.0	-.88	.45
41	low–pass 3400	8kHz/poly	32.3	17.4	-.85	.47
41	low–pass 3400, μ–law	8kHz/poly	33.1	16.7	-.86	.46
41	simulated telephone	8kHz/poly	26.4	13.9	-.80	.46
8	telephone calls	8kHz/mono	32.9	12.8	-.55	.27
8	telephone calls	8kHz/poly	37.0	15.1	-.75	.32

Table 3: Evaluation results for the four different recognizers for the 41 patients and for the 8 real phone calls.

training data of the speech recognizer for the 8 kHz was downsampled close–talk data and not real telephone data. We chose this way instead of using real telephone training data, since we wanted the telephone recognizer to be trained with the same training data as the recognizer for the close–talk data. Reducing the acoustical distance of training and evaluation data might lower the loss of correlation. An acoustic comparison (see Figure 2) of the 8 kHz resampled data to the real telephone data shows that the application of a low–pass filter with a cutoff–frequency of 3800 Hz and μ–law quantization lead to a good acoustic distance. By modifying the training data accordingly, we expect more robust recognition results. Furthermore, we expect better recognition rates by modifying the feature extraction, which is our current research.

The results in Table 3 show that for a larger corpus the polyphone–based recognizer leads to better correlation with the experts’ group. Thus the results from (Schuster et al., 2005) for 18 patients, where the monophone–based recognizer showed better agreement, were not confirmed.

Experiments with the 8 real telephone calls support these conclusions, even though this database is way too small to draw conclusions. The WA for the real telephone data is higher than for the simulated calls, probably for the

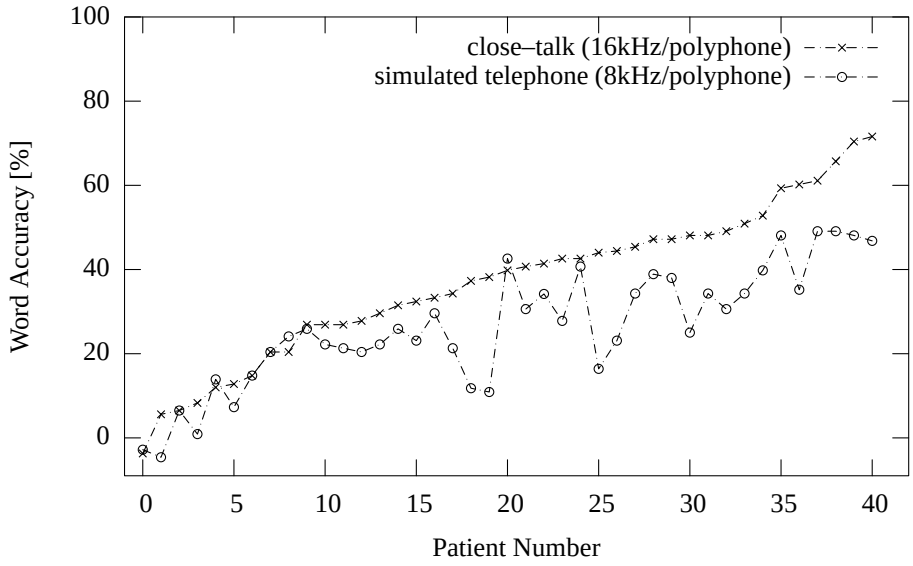

Figure 3: WAs of the 41 close–talk recordings compared to the simulated telephone recordings using polyphone–based recognizers. The recordings are ordered with increasing WA for the close–talk recordings.

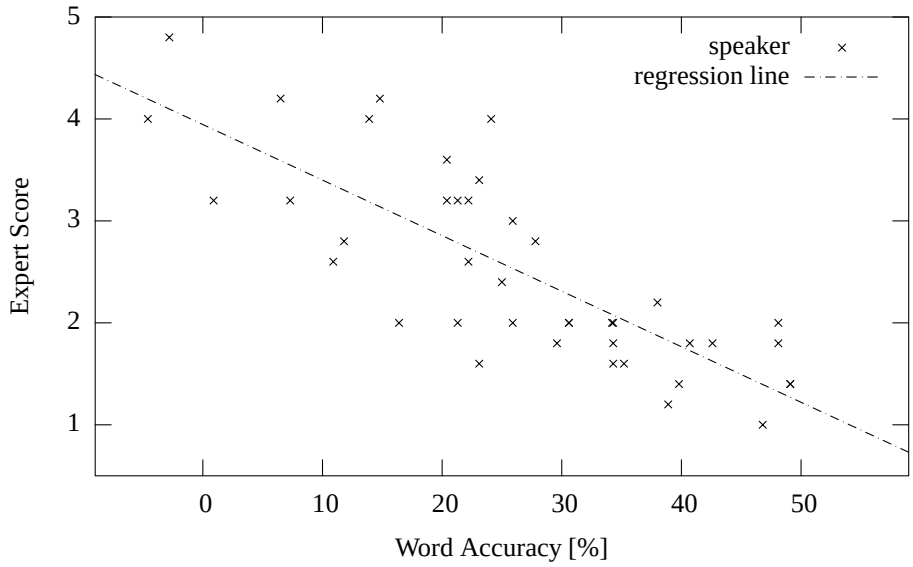

Figure 4: WA for the 41 recordings in comparison to the average of the experts' intelligibility scores using simulated telephone data and the polyphone–based recognizer.

reasons given above. The reduced κ values could be caused by the fact that the human ratings refer to a different recording and by the small corpus size. We are currently collecting a larger telephone corpus to verify the results presented in this paper.

8. Acknowledgments

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

9. References

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

D.H. Brown, F.J.M. Hilgers, J.C. Irish, and A.J.M. Balm.

2003. Postlaryngectomy Voice Rehabilitation: State of the Art at the Millennium. *World J Surg*, 27(7):824–831.

D.V. Cicchetti. 1976. Assessing inter-rater reliability for rating scales: Resolving some basic issues. *British Journal of Psychiatry*, 129(5):452–456.

M. Davies and J.L. Fleiss. 1982. Measuring agreement for multinomial data. *Biometrics*, 38(4):1047–1051.

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

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

J. Robbins, H.B. Fisher, E.C. Blom, and M.I. Singer. 1984. A Comparative Acoustic Study of Normal, Esophageal, and Tracheoesophageal Speech Production. *Journal of Speech and Hearing Disorders*, 49:202–210.

E. G. Schukat-Talamazzini. 1995. *Automatische Spracherkennung – Grundlagen, statistische Modelle und effiziente Algorithmen*. Vieweg, Braunschweig.

M. Schuster, E. Nöth, T. Haderlein, S. Steidl, A. Batliner, and F. Rosanowski. 2005. Can you Understand him? Let's Look at his Word Accuracy — Automatic Evaluation of Tracheoesophageal Speech. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, USA. IEEE Computer Society Press.

M. Schuster, T. Haderlein, E. Nöth, J. Lohscheller, U. Eysholdt, and F. Rosanowski. 2006. Intelligibility of Laryngectomees' Substitute Speech: Automatic Speech Recognition and Subjective Rating. *European Archives of Oto-Rhino-Larngology and Head & Neck*, 263:188–193.

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

G. Stemmer. 2005. *Modeling Variability in Speech Recognition*, volume 19 of *Studien zur Mustererkennung*. Logos Verlag, Berlin.

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