Original Paper

Folia Phoniatrica et Logopaedica

Folia Phoniatr Logop 2009;61:12–17
DOI: 10.1159/000187620

Published online: December 20, 2008

# Application of Automatic Speech Recognition to Quantitative Assessment of Tracheoesophageal Speech with Different Signal Quality

Tino Haderlein[a, b] Korbinian Riedhammer[b] Elmar Nöth[b] Hikmet Toy[a]
Maria Schuster[a] Ulrich Eysholdt[a] Joachim Hornegger[b] Frank Rosanowski[a]

[a]Department of Phoniatrics and Pedaudiology, and [b]Chair of Pattern Recognition (Computer Science 5), University of Erlangen-Nuremberg, Erlangen, Germany

**Key Words**
Laryngectomy · Substitute speech · Automatic speech recognition · Agreement measures

**Abstract**
***Objective:*** Tracheoesophageal voice is state-of-the-art in voice rehabilitation after laryngectomy. Intelligibility on a telephone is an important evaluation criterion as it is a crucial part of social life. An objective measure of intelligibility when talking on a telephone is desirable in the field of postlaryngectomy speech therapy and its evaluation. ***Patients and Methods:*** Based upon successful earlier studies with broadband speech, an automatic speech recognition (ASR) system was applied to 41 recordings of postlaryngectomy patients. Recordings were available in different signal qualities; quality was the crucial criterion for this study. ***Results:*** Compared to the intelligibility rating of 5 human experts, the ASR system had a correlation coefficient of $r = -0.87$ and Krippendorff's $\alpha$ of 0.65 when broadband speech was processed. The rater group alone achieved $\alpha = 0.66$. With the test recordings in telephone quality, the system reached $r = -0.79$ and $\alpha = 0.67$. ***Conclusion:*** For medical purposes, a comprehensive diagnostic approach to (substitute) voice has to cover both subjective and objective tests. An automatic recognition system such as the one proposed in this study can be used for objective intelligibility rating with results comparable to those of human experts. This holds for broadband speech as well as for automatic evaluation via telephone.



## Introduction

Laryngectomy for laryngeal or hypopharyngeal cancer affects many aspects of life [1] with loss of ability of vocal communication being an outstanding stigma for the affected persons [2]. Tracheoesophageal (TE) substitute voice is state-of-the-art voice rehabilitation after laryngectomy [3]. In comparison to normal voices, the quality of substitute voices is 'low', e.g. the change of pitch and volume is limited, and intercycle frequency perturbations result in a hoarse voice and reduced intelligibility [4].

In order to improve postlaryngectomy speech therapy, an objective means of rating intelligibility is needed as it might improve the assessment in clinical routine with respect to evidence-based medicine. At present, there is no consensus on which measures are best to evaluate postlaryngectomy voice rehabilitation. In a previous work, we

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com


1021–7762/09/0611–0012$26.00/0

Accessible online at:
www.karger.com/fpl

Tino Haderlein
Abteilung für Phoniatrie und Pädaudiologie, Bohlenplatz 21
DE–91054 Erlangen (Germany)
Tel. +49 9131 852 7872, Fax +49 9131 303 811
E-Mail Tino.Haderlein@informatik.uni-erlangen.de

showed that an automatic speech recognition (ASR) system can be used to rate intelligibility in broadband speech of postlaryngectomy speakers [5]. The ASR system 'understands' different words and hence is a suitable basis for automatic evaluation of intelligibility.

Communication may not only be affected by the persons involved, but also by the transmission channel between them: the telephone is a crucial part of patients' social life, and especially in emergency cases it is necessary for laryngectomees to have a means of communication that does not require them to leave their home. For this reason, intelligibility on a telephone reflects everyday communication that is important for the patient. In this paper, we examine how well TE telephone speech is processed by an ASR system and how well the objective measure of intelligibility computed from the recognition results corresponds to perceptive evaluation. The listeners were voice professionals since the special interest of the study was on automatic support methods for clinical purposes.

The crucial criterion for this study is signal quality. The test persons were recorded with a close-talk microphone only. We produced data that differed from these broadband samples only in signal quality by replaying the broadband recordings using a loudspeaker and recording them again through a telephone line. In this way, we excluded all other possible influence factors except signal quality.

## Materials and Methods

*Subjects and Test Samples*

Forty-one laryngectomees (2 female and 39 male) with TE substitute voice, on the average 62.0 years old (SD 7.7 years), were evaluated. The 2 women were 54.4 and 70.8 years old and did not have a significant influence on the age distribution of the group. Each speaker read the German text 'Der Nordwind und die Sonne'. It is a phonetically balanced text consisting of 108 words (71 disjoint) which is used in speech therapy in German-speaking countries; the English version is known as 'The North Wind and the Sun' [6]. 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).

We created three additional versions of these broadband data:

*Low-Pass 3,400.* The recordings were down-sampled with a sampling frequency of 8 kHz, and a low-pass filter with a cutoff frequency of 3,400 Hz was applied because this is also done during telephone transmission.

*Low-Pass 3,400, μ-Law.* In order to simulate the loss due to the logarithmic encoding in the telephone channel, we converted the linearly quantized signals to μ-law companded signals [7] and back to linearly quantized signals. The samples were recorded with 8 kHz and quantized with 16 bit (linear).

*Simulated Telephone.* In order to get a 'telephone quality' version of the signals, we played back the broadband recordings using a standard PC and loudspeaker (quadral SAM38) in a quiet office environment and placed a telephone headset in front of the loudspeaker. The replayed sound files were recorded with an automatic dialogue system over the telephone with 8 kHz and 16 bit linear.

*Recognition System*

The speech recognition system used for the experiments was developed at the Chair of Pattern Recognition at the University of Erlangen-Nuremberg [8, 9]. A commercial version of the system is used in high-end telephony-based conversational dialogue systems by Sympalog Voice Solutions.

The system is based on semicontinuous Hidden Markov Models, which define a statistical model for each different phoneme to be recognized. This is a standard method in automatic speech recognition [10]. When one model is trained for each phoneme, the recognizer is called 'monophone-based'. Context-dependent phoneme models were also applied. They take into account coarticulation effects and train e.g. different models for the core phone [I] in the phone context v[I]n (as in *win* or *winning*) or k[I]d (as in *kid*, *kidney*, etc.). We use a special kind of these models where the context that is chosen can be of arbitrary length. They are referred to as 'polyphone' models [11]. For this study, polyphone-based and monophone-based recognizers were compared in order to determine which approach yields better results.

The recordings are segmented to form 16-ms 'frames' at a frame shift rate of 10 ms. In each of these frames, one core phone was classified. The recognized phones were combined to words according to the list of the words in the text 'Der Nordwind und die Sonne'. More technical details on this method were published in Gallwitz [8] and Stemmer [9].

The output of the system was the recognized word sequence from which word accuracy was obtained. It was computed from the comparison between the recognized sequence and the reference text consisting of $n_{all} = 108$ words. With the number of words that were wrongly substituted ($n_{sub}$), deleted ($n_{del}$) and inserted ($n_{ins}$) by the recognizer, word accuracy (WA) in percent is given as

$$WA = [1 - (n_{sub} + n_{del} + n_{ins})/n_{all}] \cdot 100.$$

In order to reduce the computational complexity in the recognition phase, we used a so-called unigram language model to weight the probability of appearance of each word model [12]. Hence, the frequency of occurrence for each single word in the text 'Der Nordwind und die Sonne' was known to the recognizer. Statistical information about sequences of words was not used because it would have corrected too many recognition errors. Word accuracy would become useless as a measure of intelligibility, because words that could not be 'understood' correctly by the system due to low voice quality would be corrected according to the statistical knowledge about the occurrence of these words.

*Training of the Acoustic Phoneme Models*

The basic training set for the acoustic phone models of the recognizer were broadband recordings from the Verbmobil project [13]. The data were recorded with a close-talk microphone at a sampling frequency of 16 kHz and quantized with 16 bit (linear). The training speakers were from all over Germany; they were

**Table 1.** Intervals for mapping the word accuracy (WA) of the speech recognition system to scores on the integer rating scale on which the expert listeners rated intelligibility

| | WA interval | | | | |
|---|---|---|---|---|---|
| | [–●●; 0] | [0; 15] | [15; 25] | [25; 40] | [40; 100] |
| Score | 5 | 4 | 3 | 2 | 1 |

asked to speak 'standard' German. All persons were normal speakers. In this way, a normal voice was defined as the reference for automatic evaluation.

In order to extend the evaluation to telephone speech, we created four different recognizers. In addition to the original system that processes recordings made with 16 kHz sampling frequency, we reduced the sampling rate of the training data to 8 kHz and applied the same low-pass filter (3,400 Hz) as for the test data. For the 16- and 8-kHz training data, we created both a polyphone-based and a monophone-based recognizer (table 3, 16 kHz/mono, 8 kHz/mono, 16 kHz/poly, 8 kHz/poly). The recognition vocabulary was reduced to the words occurring in the text 'Der Nordwind und die Sonne'.

The training set of the recognizer for the 8-kHz data consisted of down-sampled broadband speech and not real telephone data. We chose this way instead of using real telephone speech for training since we wanted the telephone recognizer to be trained with the same recordings as the recognizer for the broadband data, just with another signal quality. This was done in order to minimize the factors having an influence on the recognition result.

*Subjective Evaluation*

A group of 5 voice professionals subjectively estimated the intelligibility of the patients while listening to a playback of the broadband recordings. A five-point Likert scale was applied to rate the intelligibility of each recording, i.e. the listeners were asked to mark one of the grades 'very high', 'rather high', 'medium', 'rather low', or 'very low'. For computation, these grades were converted to integer values from 1 (very high intelligibility) to 5. In this manner an averaged mark, expressed as a floating point value, could be calculated for each patient by averaging the scores of the single raters. Since the difference between the broadband and the other recordings was only the signal quality, the expert raters' evaluation of the broadband data was used as reference for all experiments. Multiple evaluations would have introduced another source of error – intra-rater discrepancy – which would have biased the results of the study.

*Measures for Inter-Rater Agreement*

For the purpose of measuring the correlation in the human experts' intelligibility ratings and the correlation between human and machine, we applied different measures. In order to compare an arbitrary number of raters, we use the weighted multi-rater $\kappa$ by Davies and Fleiss [14], an extension of Cohen's $\kappa$. Two raters' scores x and y are weighted as proposed by Cicchetti [15] with

$$w(x, y) = 1 - |(x - y)/(C - 1)|$$

where C is the number of different rating grades (here: 5). The minimum value of $\kappa$ is below 0 and depends on the input data, the maximum of 1 means perfect agreement. However, $\kappa$ may show unexpected behavior when a rater does not use the entire possible range for his or her rating, so $\kappa$ may be low even if the level of agreement is high [16–18]. Therefore, the often mentioned intervals of a 'moderate' agreement for $0.4 \leq \kappa \leq 0.75$ and a 'good' agreement for $\kappa > 0.75$ are inadequate [19]. For reliable computation of $\kappa$, it is necessary for every rater to choose each one of the possible categories at least once. In many studies this is not fulfilled, and it is also not the case with our data. A more severe problem is when one rater does not give a judgment at all for some of the test data. These data usually have to be excluded from further evaluation.

A less known measure that is able to cope with these commonly occurring problems is Krippendorff's $\alpha$ [20]. It can be applied to data described by any metric or level of measurement and also to incomplete or missing data. The minimum value that $\alpha$ can get is 0 in the case that the agreement observed was just a product of chance, not of the raters' competence. When the raters agree perfectly, then $\alpha$ reaches its maximum at 1. Inter-rater reliability is usually regarded as being sufficient if $\alpha$ is greater than approximately 0.70 [20]. In the case of integer scales with a numerical order, the distance metric that $\alpha$ is based on will be chosen of the interval metric type, i.e. the difference in the data values is expressed by

$$\delta^2(x, y) = (x - y)^2.$$

According to their definition, $\kappa$ and $\alpha$ were computed for the rater group as a whole by averaging the respective measures of all possible rater pairs. For the agreement between the expert listeners and the speech recognizers, the respective recognizer was added as a 5th expert to all possible groups of 4 experts. Then inter-rater agreement among these groups was determined and averaged. However, $\kappa$ and $\alpha$ are only defined for integer input values. Since the word accuracy value provided by the speech recognizers is a float range number, it had to be mapped to a five-point scale to fit the range of human evaluation. The intervals of word accuracy values that were mapped to the specific numbers are subsumed in table 1. The interval boundaries are based upon earlier findings of a study with broadband recordings of 18 TE speakers [21]. The determined intervals had minimized the difference between expert ratings and the converted word accuracy values.

In order to compare the weighted $\kappa$ and $\alpha$ to more introduced measures, we also computed Pearson's product-moment correlation coefficient r and Spearman's rank-order correlation coefficient $\rho$ for the average rater against the automatic evaluation results. To judge the agreement between the different raters, we calculated the correlation between each rater's 'intelligibility' rating and the average of the 4 other raters. Due to the range of the average number, it was not possible to determine $\kappa$ and $\alpha$. Note that Pearson's r has the disadvantage of standardizing the input values and measuring covariation only. It also does not reflect the portion of agreement that may occur by pure coincidence.

## Results

Table 2 shows the correlation r and ρ between each rater and the average of the 4 remaining raters, and the average correlation coefficient of all these single results. The weighted multi-rater κ for the group of the 5 raters was 0.45; Krippendorff's α reached 0.66.

In table 3, the recognition and correlation results of the speech recognizers for the 41 patients are given. The correlation between the word accuracy of the respective polyphone-based recognizers and the average of the expert intelligibility scores is reduced from –0.87 to –0.79 when evaluating simulated telephone calls instead of broadband speech. For the monophone-based recognizers, the correlation dropped from –0.82 to –0.68. The coefficient is negative because high recognition rates came from 'good' voices with a low score number and vice versa. The μ-law coding of the telephone transmission obviously does not have a deteriorating effect on the speech signals; the differences between the results on linear and μ-law coding are not significant ($p > 0.1$).

In contrast to the correlation coefficients r and ρ, the values of κ and α are not significantly different ($p > 0.1$) for the poorer acoustic telephone quality. For the polyphone-based recognizer, even a slight increase could be observed for the 8-kHz data in comparison to the original broadbanding data. This is not only in contrast to the r values, but also to overall word accuracy. The polyphone-based recognizers agree better with the group of experts than the monophone-based systems.

## Discussion

Until now, no generally accepted objective method for the evaluation of speech restoration outcome after laryngectomy has been available. Here, we present an automatic objective measurement of the clinically valid intelligibility criterion based upon the word accuracy of an automatic speech recognition system. It is achieved by the analysis of running speech rather than sustained vowels like other approaches for measuring laryngeal voice quality [22–24]. Hence, it is possible to evaluate the patient's voice and speech at the same time. A similar approach was introduced by Moerman et al. [25], but the text used there was shorter: it contained 18 words only. Correlations to human ratings were only given for the 'overall

**Table 2.** Correlation (Pearson's r and Spearman's ρ) between single raters' intelligibility scores and the average of the 4 other raters

| | Rater | | | | | |
|---|---|---|---|---|---|---|
| | K | L | R | S | U | avg. |
| r | 0.82 | 0.80 | 0.81 | 0.85 | 0.77 | 0.81 |
| ρ | 0.82 | 0.81 | 0.76 | 0.83 | 0.76 | 0.79 |

avg. = Average over all raters. The weighted multi-rater κ for the group of the 5 raters was 0.45, Krippendorff's α reached 0.66.

**Table 3.** Agreement (Pearson's r, Spearman's ρ, multi-rater κ [14], and Krippendorff's α) between human intelligibility ratings and automatically computed word accuracy (WA) for TE speech recordings of different signal quality

| Recording | Data/recognizer | WA | | r | ρ | κ | α |
|---|---|---|---|---|---|---|---|
| | | avg. | SD | | | | |
| Broadband | 16 kHz/mono | 35.3 | 13.7 | –0.82 | –0.82 | 0.41 | 0.60 |
| Low-pass 3,400 | 8 kHz/mono | 33.4 | 12.1 | –0.80 | –0.76 | 0.42 | 0.62 |
| Low-pass 3,400, μ-law | 8 kHz/mono | 33.6 | 12.7 | –0.77 | –0.75 | 0.42 | 0.62 |
| Simulated telephone | 8 kHz/mono | 28.4 | 10.3 | –0.68 | –0.69 | 0.41 | 0.60 |
| Broadband | 16 kHz/poly | 36.9 | 18.0 | –0.87 | –0.85 | 0.44 | 0.65 |
| Low-pass 3,400 | 8 kHz/poly | 32.3 | 17.4 | –0.84 | –0.84 | 0.47 | 0.68 |
| Low-pass 3,400, μ-law | 8 kHz/poly | 33.1 | 16.7 | –0.85 | –0.85 | 0.46 | 0.66 |
| Simulated telephone | 8 kHz/poly | 26.4 | 13.9 | –0.79 | –0.80 | 0.46 | 0.67 |

Four different recognizers (16 or 8 kHz sampling frequency and monophone-based or polyphone-based, respectively) were applied.

impression' of the substitute voice, so no direct comparison with our study is possible.

Another aspect that is different to most other studies is that we also examined telephone speech. A study on the classification of sustained vowels [24] used the same type of simulated telephone data as our experiments, i.e. the broadband recordings were played back by a loudspeaker and transmitted via telephone. The automatic classification into 'normal' and 'pathologic' showed a success rate of 74.2% while it was beyond 90% for broadband speech. The main difference in the data was, like in our study, that the frequency region above 4 kHz was cut off by the telephone transmission. Obviously, these frequencies are important for the acoustic distinction between a normal and a pathologic voice and hence the success rate for telephone data is lower. The results of our study indicate, however, that the automatic evaluation of the criterion 'speech intelligibility' of TE speakers is not severely affected by the different signal quality. It has to be noted, however, that there were much more test data for each patient. Nevertheless, overall word accuracy for the simulated telephone calls was strongly reduced 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 recognition rates of the simulated telephone data are expected to be a lower bound for the recognition rates for real telephone calls. The 2 female speakers in the test set did not have a significant influence on the recognition rates since the recognizers were trained with both male and female speech.

For this study, patients read a standard text, and voice professionals evaluated intelligibility. It is often argued that intelligibility should be evaluated by an 'inverse intelligibility test': the patient utters a subset of words and sentences from a carefully built corpus. A naïve listener writes down what he or she heard. The percentage of correctly understood words is a measure of the intelligibility of the patient. In a study on one of these standardized tests, the German Post-Laryngectomy Telephone Test (PLTT) [26], however, we showed that one naïve rater alone is not enough to achieve reliable results because inter-rater correlation with other raters is too low [27]. Our intention is also to design an automatic support for speech therapy, so the reference data have to be obtained from trained listeners first.

The reason why we use a standard text is the following: when automatic speech evaluation is performed for instance with respect to prosodic phenomena, e.g. word durations or percentage of voiced segments [28], then comparable results for all patients can only be achieved when all the patients read the same defined words or text. This means that an inverse intelligibility test can no longer be performed, and intelligibility has to be rated on a grading scale instead. However, inter-rater correlation between an objective, automatic evaluation method and a group of experts rating intelligibility on a five-point scale was well above 0.8 [5, 21, 29]. For an objective, automatic version of the PLTT, correlations between the average naïve listener and the automatic results were in the same range [27]. Hence, the text-based evaluation performed by trained listeners is as reliable as the inverse intelligibility test with the naïve raters.

The conclusion from our experiments is that automatic evaluation of the intelligibility of substitute voices is not only possible on broadband speech, but also on telephone speech. The degree of agreement between automatic and human evaluation is not worse than within a group of human raters. For $\alpha$, some experiments on automatic evaluation are just under 0.70, which is regarded as the lower threshold for reliable agreement. But even human raters among themselves reach only $\alpha = 0.66$, although they are all experienced. For automatic evaluation, polyphone-based recognizers based on context-dependent phoneme models appear superior to monophone-based systems. Although the agreement between human and machine is already on the same level as among human raters, we expect even better results by improved automatic word recognition in the future.


## Acknowledgment

This work was funded by the German Cancer Aid (Deutsche Krebshilfe, grant 106266).


## References


1 Schuster M, Lohscheller J, Kummer P, Hoppe U, Eysholdt U, Rosanowski F: Quality of life in laryngectomees after prosthetic voice restoration. Folia Phoniatr Logop 2003;55:211–219.

2 Devins GM, Stam HJ, Koopmans JP: Psychosocial impact of laryngectomy mediated by perceived stigma and illness intrusiveness. Can J Psychiatry 1994;39:608–616.

3 Brown DH, Hilgers FJM, Irish JC, Balm AJM: Postlaryngectomy voice rehabilitation: state of the art at the millennium. World J Surg 2003;27:824–831.

4 Schutte HK, Nieboer GJ: Aerodynamics of esophageal voice production with and without a Groningen voice prosthesis. Folia Phoniatr Logop 2002;54:8–18.

5 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 2006;263:188–193.
6 International Phonetic Association: Handbook of the International Phonetic Association. Cambridge, Cambridge University Press, 1999.
7 Telecommunication Standardization Sector of the ITU (ITU-T): Recommendation G.711: Pulse code modulation (PCM) of voice frequencies. Geneva, 1988.
8 Gallwitz F: Integrated Stochastic Models for Spontaneous Speech Recognition; in Studien zur Mustererkennung. Berlin, Logos, 2002, vol 6.
9 Stemmer G: Modeling Variability in Speech Recognition; in Studien zur Mustererkennung. Berlin, Logos, 2005, vol 19.
10 Huang X, Acero A, Hon HW: Spoken Language Processing. Upper Saddle River, Prentice Hall, 2001.
11 Schukat-Talamazzini EG, Niemann H, Eckert W, Kuhn T, Rieck S: Automatic speech recognition without phonemes; in Proc Eur Conf on Speech Commun and Technol (Eurospeech). Berlin, European Speech Communication Association (ESCA), 1993, pp 129–132.
12 Jelinek F: Self-organized language modeling for speech recognition; in Waibel A, Lee KF (eds): Readings in Speech Recognition. San Mateo, Morgan Kaufman, 1990, pp 450–506.
13 Wahlster W (ed): Verbmobil: Foundations of Speech-to-Speech Translation. Berlin, Springer, 2000.
14 Davies M, Fleiss J: Measuring agreement for multinomial data. Biometrics 1982;38:1047–1051.
15 Cicchetti D: Assessing inter-rater reliability for rating scales: resolving some basic issues. Br J Psychiatry 1976;129:452–456.
16 Uebersax J: Diversity of decision-making models and the measurement of interrater agreement. Psychol Bull 1987;101:140–146.
17 Feinstein A, Cicchetti D: High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol 1990;43:543–549.
18 Cicchetti D, Feinstein A: High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol 1990;43:551–558.
19 Fleiss J: Statistical Methods for Rates and Proportions, ed 2. New York, Wiley, 1981.
20 Krippendorff K: Content Analysis, an Introduction to Its Methodology, ed 2. Thousand Oaks, Sage, 2003.
21 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 7th Int Conf Text, Speech and Dialogue (TSD 2004). Lecture Notes in Artificial Intelligence, vol 3206. Berlin, Springer, 2004, pp 331–338.
22 Fröhlich M, Michaelis D, Strube HW, Kruse E: Acoustic voice analysis by means of the hoarseness diagram. J Speech Lang Hear Res 2000;43:706–720.
23 van Gogh CDL, Festen JM, Verdonck-de Leeuw IM, Parker AJ, Traissac L, Cheesman AD, Mahieu HF: Acoustical analysis of tracheoesophageal voice. Speech Commun 2005;47:160–168.
24 Moran RJ, Reilly RB, de Chazal P, Lacy PD: Telephony-based voice pathology assessment using automatic speech analysis. IEEE Trans Biomed Eng 2006;53:468–477.
25 Moerman M, Pieters G, Martens JP, van der Borgt MJ, Dejonckere P: Objective evaluation of the quality of substitution voices. Eur Arch Otorhinolaryngol 2004;261:541–547.
26 Zenner HP: The Postlaryngectomy Telephone Intelligibility Test (PLTT); in Herrmann IF (ed): Speech Restoration via Voice Prosthesis. Berlin, Springer, 1986, pp 148–152.
27 Haderlein T, Riedhammer K, Maier A, Nöth E, Toy H, Rosanowski F: An automatic version of the Post-Laryngectomy Telephone Test; in Matoušek V, Mautner P (eds): Proc 10th Int Conf Text, Speech and Dialogue (TSD 2007). Lecture Notes in Artificial Intelligence, vol 4629. Berlin, Springer, 2007, pp 238–245.
28 Haderlein T, Nöth E, Toy H, Batliner A, Schuster M, Eysholdt U, Hornegger J, Rosanowski F: Automatic evaluation of prosodic features of tracheoesophageal substitute voice. Eur Arch Otorhinolaryngol 2007;264:1315–1321.
29 Maier A, Haderlein T, Schuster M, Nkenke E, Nöth E: Intelligibility is more than a single Word: Quantification of Speech Intelligibilty by ASR and Prosody; in Matoušek V, Mautner P (eds): Proc 10th Int Conf Text, Speech and Dialogue (TSD 2007). Lecture Notes in Artificial Intelligence, vol 4629. Berlin, Springer, 2007, pp 278–285.