# Automatic intelligibility assessment of pathologic speech over the telephone

Tino Haderlein, Elmar Nöth, Anton Batliner, Ulrich Eysholdt, Frank Rosanowski

**Angaben zur Veröffentlichung / Publication details:**

Haderlein, Tino, Elmar Nöth, Anton Batliner, Ulrich Eysholdt, and Frank Rosanowski. 2011. "Automatic intelligibility assessment of pathologic speech over the telephone." *Logopedics Phoniatrics Vocology* 36 (4): 175–81. https://doi.org/10.3109/14015439.2011.607470.

# Automatic Intelligibility Assessment of Pathologic Speech over the Telephone

Tino Haderlein[1,2], Elmar Nöth[1], Anton Batliner[1], Ulrich Eysholdt[2], Frank Rosanowski[2]

[1]Pattern Recognition Lab (Computer Science 5), University of Erlangen-Nuremberg
Martensstraße 3
91058 Erlangen
Germany

[2]Department of Phoniatrics and Pediatric Audiology, University Hospital Erlangen
Bohlenplatz 21
91054 Erlangen
Germany

Running Title: Automatic Intelligibility Rating on the Telephone

Corresponding author:
Tino Haderlein
Lehrstuhl für Mustererkennung (Informatik 5)
Friedrich-Alexander-Universität Erlangen-Nürnberg
Martensstraße 3
91058 Erlangen
Germany
E-Mail: Tino.Haderlein@informatik.uni-erlangen.de
Phone: +49 9131 852-7872
Fax: +49 9131 303811

## Abstract

Tino Haderlein, Elmar Nöth, Anton Batliner, Ulrich Eysholdt, Frank Rosanowski

Automatic Intelligibility Assessment of Pathologic Speech over the Telephone

Objective assessment of intelligibility on the telephone is desirable for voice and speech assessment and rehabilitation. 82 patients after partial laryngectomy read a standardized text which was synchronously recorded by a headset and via telephone. Five experienced raters assessed intelligibility perceptually on a 5-point scale. Objective evaluation was performed by Support Vector Regression on the word accuracy (WA) and word correctness (WR) of a speech recognition system, and a set of prosodic features. WA and WR alone exhibited correlations to human evaluation between |r|=0.57 and |r|=0.75. The correlation was r=0.79 for headset and r=0.86 for telephone recordings when prosodic features and WR were combined. The best feature subset was optimal for both signal qualities. It consists of WR, the average duration of the silent pauses before a word, the standard deviation of the fundamental frequency on the entire sample, the standard deviation of jitter, and the ratio of the durations of the voiced sections and the entire recording.

**Keywords** Laryngectomy, Telephone Speech, Automatic Speech Recognition, Prosodic Analysis, Intelligibility Assessment

## Introduction

Perceptual voice and speech evaluation for clinical and scientific purposes is biased and time-consuming. Automatically computed, objective measures help to reduce costs, and the problem of inter- and intra-rater variability is eliminated, because an automated evaluation algorithm always yields the same result for one specific speech recording. In this way, it can be used as objective assessment method during voice and speech rehabilitation therapy. This article introduces such a method which is independent of a particular therapist's experience.

Although a basic protocol for functional assessment of voice pathology was established (1), there is still no final decision about which automatic methods should be applied for this purpose. Currently available software usually evaluates isolated voice properties but not speech aspects (2,3,4). However, the necessity for the analysis of more complex speech elements than vowels, especially for criteria like speech intelligibility or prosodic aspects, has been pointed out in the literature (5,6,7,8,9). Intelligibility may not only be affected by the communication partners but also by the transmission channel between them: The telephone is a crucial part of modern communication society, especially for the social life of elderly people whose physical abilities are restricted. For this reason, intelligibility on a telephone is an aspect of everyday communication which is already assessed in self-evaluation questionnaires on voice-related quality of life (10).

In order to rate intelligibility of pathologic speakers, automatic speech recognition (ASR) systems can be applied (11,12,13,14). Additionally, prosodic analysis is widely used in automatic speech analysis on normal voices (15,16,17,18,19). However, it was also proved applicable to evaluate voice and speech disorders (20,21). Even for telephone speech of partially laryngectomized persons, prosodic measures were applied (11). The combination of a speech recognition system and a "prosody module", however, has not been tested until now and will be presented in this article.

The goal of this study was the identification of a set of measures from speech recognition and prosodic analysis which allows evaluation of intelligibility both on high-quality speech recordings and telephone recordings. This set was supposed to be applicable to a broad range of voice pathology degrees. For this reason, the study was performed with partially laryngectomized persons.

## Material and Methods

### Patient Group

82 persons (68 men and 14 women) after partial laryngectomy were involved in this study. Their average age was 62.3SD8.8 years; the youngest speaker was 41.1, the oldest one was 86.1 years old. In most cases, only a small tumor had been removed (T1: N=48, T2: N=19, T3: N=11, T4: N=4). At the time of investigation, none of the persons suffered from recurrent tumor growth or metastases. Informed consent had been obtained by all participants prior to the examination. The study respected the principles of the World Medical Association (WMA) Declaration of Helsinki on ethical principles for medical research involving human subjects (22) and has been approved by the ethics committee of the University of Erlangen-Nuremberg. The participants read the German version of the tale "The North Wind and the Sun" (23) which is widely used in medical speech evaluation in German-speaking and other countries. It consists of 71 disjoint words and 108 words in total (172 syllables).

The patients were recorded simultaneously by a close-talk microphone (Logitech Premium Stereo Headset 980369-0914) and via a landline telephone. For the close-talk samples, 16 kHz sampling frequency and 16 bit linear amplitude resolution were applied. This is standard in many applications using automatic speech recognition. It captures all important frequencies and time-related phenomena occurring in human speech. The telephone samples, which underwent the usual a-law companding used in telephone transmission, were afterward resampled to 8 kHz and 16 bit.

Perceptual Evaluation

Five experienced voice professionals evaluated the intelligibility of each recording. The speech samples were played to the experts once via loudspeakers in a quiet seminar room without disturbing noise or echoes. Forty-four recordings of the headset and the telephone data were evaluated a second time after four weeks in order to compute intra-rater correlation. Rating was performed on a five-point Likert scale. For computation of average scores for each patient, these grades were converted to integer values (1 = “very high”, 2 = “rather high”, 3 = “medium”, 4 = “rather low”, 5 = “very low”).

Recognition System

The speech recognition system used for the experiments (24) is based on semi-continuous Hidden Markov Models (HMM) which define a statistical model for each different phoneme to be recognized. The basic acoustic measures (features) for the recognition are Mel-Frequency Cepstrum Coefficients which are also used for voice assessment (25). For speech recognition with our system, they are computed for each section of 16 ms duration. Their purpose is to extract those patterns which are characteristic for a certain phoneme but speaker-independent from the sample. Many features representing the same phoneme form a cluster in the multi-dimensional feature space. The position of this cluster is described by a weighted sum of Gaussians which is determined during the training phase of the system. Hence, the parameters of those Gaussians define the model of a phoneme. During the recognition phase, it is tested for each feature vector of a recording to which cluster and phoneme it belongs.

The phoneme models are context-dependent. 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 special “polyphone” models (26) where the context can be chosen arbitrarily large. The basic training set for the acoustic phone models for this study were broadband recordings. The 578 training speakers (304 male, 274 female) were normal speakers from all over Germany. In this way a normal voice was defined as the reference for automatic evaluation. The average age of these persons was 27 years. About 80% of them were between 20 and 29 years old, less than 10% were over 40. In order to extend the evaluation to telephone speech, a second recognizer was trained. We resampled the original training data with 8 kHz and applied a band-pass filter (300 to 3400 Hz). This simulates telephone speech quality.

For speech recognition, the recognized phones are combined to words according to the list of the words in “The North Wind and the Sun” (24). The word accuracy (WA) and the word correctness (WR) are obtained from the comparison between the recognized word 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 system, the word accuracy is given as

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

The word correctness omits the wrongly inserted words:

$$WR\ [\%] = [1 - (n_{sub} + n_{del})/n_{all}] \cdot 100 \qquad (2)$$

Although both measures are usually given in percent, a high $n_{ins}$ can cause the WA to become negative. This happens only for very severely distorted voices or recordings when the recognizer identifies a large number of short words where actually longer words were uttered. In the data for this study, this case did not occur. The negative values do not restrict the usability of the data for statistical processing in any way.

In order to reduce the computational complexity in the recognition phase, a "language model" of possible speech input is usually added as another source of information. It contains probabilities about word sequences occurring in natural language. Hence, many errors from the pure acoustic recognition phase can be eliminated. However, for the purpose of automatic assessment of intelligibility, this is a disadvantage. The more errors are corrected by using linguistic knowledge, the worse match human and automatic evaluation (12). This makes WA and WR useless as measures for intelligibility. For this reason, our recognizer did not have information about word pairs, triples, or longer phrases. It used only a unigram language model instead, i.e. the frequency of occurrence of single words in the text reference was known to the recognizer.

Prosodic Features

For each word provided by the speech recognizer, a vector of prosodic features is computed by the prosody module. There are three basic groups of features. Duration features represent word and pause durations. Energy features contain information about maximum and minimum energy, their respective positions in the word, the energy regression coefficient and mean square error. Similarly the $F_0$ features, based on the detected fundamental frequency, comprise information about the extremal $F_0$ values and their positions, voice onset and offset with their positions, and also the regression coefficient and mean square error of the $F_0$ trajectory. Duration, energy, and $F_0$ values are stored as absolute and normalized values. The 28 basic features are computed in different contexts, i.e. in intervals containing the single word or pause only or a word-pause-word interval. In this way, 95 features are computed for each word (16,21,27).

Besides the 95 local features per word, 16 "global" features are computed for intervals of 15 words length each. They are derived from jitter (fluctuations of $F_0$), shimmer (fluctuations of intensity), and the number of detected voiced and unvoiced sections in the speech signal (16). They cover the means and standard deviations of jitter and shimmer, the number, length and maximum length of voiced and unvoiced sections, the ratio of the numbers of voiced and unvoiced sections, the ratio of the length of the voiced sections to the length of the signal, and the same for unvoiced sections. The standard deviation of the $F_0$ was measured in two ways: It was computed over the voiced sections and also over all sections of the speech recordings. In the latter case, each unvoiced frame contributed a value of 0. Hence, it incorporated also information about the percentage of frames where no regular voice signal was detected. Since all patients read the same text, this was supposed to indicate the degree of pathology.

The speech experts gave ratings for the entire text. In order to receive also one single value for each feature that could be compared to the human ratings, the average of each prosodic feature over the entire recording served as a final feature value.

Support Vector Regression

A Support Vector Machine (SVM) performs a binary classification based on a hyperplane separation between two class areas in a multi-dimensional feature space. SVMs can also be used for Support Vector Regression (SVR, 28). The general idea of regression is to use the element vectors of the training set to approximate a function which tries to predict the target value of a given vector of the test set. For this study, the sequential minimal optimization algorithm (SMO, 28) of the Weka toolbox (29) was applied for this purpose. The automatically computed measures WA, WR, and all prosodic features served as the training set for the regression, and the test set consisted of the perceptually assessed intelligibility scores.

In order to find the best subset of WA, WR, and the prosodic features to model the subjective ratings, a correlation-based feature selection method (20, p.59-61) was applied. It was performed separately for close-talk and telephone speech in a 10-fold cross-validation manner. The features with the highest ranks were then used as input for the SVR.

## Results

Intelligibility was rated lower for the telephone recordings by the speech recognizer (47.0% WA, 53.9% WR; see tables I and II) and the raters (score 3.25) than for the close-talk recordings (51.8% WA, 57.2% WR, score 2.91). The differences, however, are not significant. The differences between the single raters are also not significant. This was confirmed by the average inter-rater correlation between one rater and the four others which was r=0.84 (table III). The worst value was r=0.79 in this case. Likewise, the intra-rater correlation was high (table IV). The average was r=0.83 for both data sets. All the single raters differed only slightly from this average, except for rater R1 who exhibited the smallest value of all (r=0.73) for the close-talk recordings and the maximum of all raters (r=0.90) for the telephone data. The average correlation between the intelligibility rating of the headset and the corresponding telephone samples amounted to r=0.77 (table V). For rater R1, it was only r=0.71 while the other experts did not differ significantly from the average.

[Insert Table I about here.]
[Insert Table II about here.]
[Insert Table III about here.]
[Insert Table IV about here.]
[Insert Table V about here.]

The correlation between perceptual and automatic evaluation is shown in table VI. WA and WR alone featured correlations to human evaluation between r=-0.57 and r=-0.75. The coefficient is negative because high recognition rates came from “good” voices with a low score number and vice versa. The human-machine correlation rose significantly ($p<0.01$) to values of r=0.79 for headset and r=0.86 for telephone recording when prosodic features and WR were combined. The best subset that was identified was optimal for both signal qualities. It consists of WR, the average duration of the silent pause before a word, the standard deviation of $F_0$ on the entire sample, the standard deviation of jitter, and the ratio of the durations of the voiced sections and the entire recording. The weighting factors of the single features in the regression formula between human and machine rating are given in table VII.

[Insert Table VI about here.]
[Insert Table VII about here.]

## Discussion

### The Best Feature Set

The best feature subset for the agreement between experts and automatic evaluation (table VII) contains only five measures which are the same for close-talk and telephone speech. The contribution of the word correctness to the human-machine correlation is obvious. The more words a human listener understands, the more also the machine identifies correctly. This has been demonstrated for other pathologies before (11,12,13). The reason why WR and not WA is better here might be the smaller degree of pathology in comparison to, for instance, total laryngectomees. For that person group, the recognition results are very bad with a high percentage of wrongly inserted words. This number may be the key to the good correlation with the low perceptual intelligibility scores. In partially laryngectomized patients, this measure loses its expressiveness because recognition performs generally better. This is also the reason why WA or WR alone cannot capture the complex phenomenon of intelligibility (table VI).

Intelligibility is also related to speaking rate and voice quality since the best feature set contains duration-based features and features measuring regularity. The duration of silent pauses between words reflects the speaking rate. We assume that speakers with longer pauses have to put more effort into speaking. This may be caused by severe alterations to the voicing organs. For instance, in incomplete glottal closure, the speaker needs to breath more frequently because of the loss of air through the glottal gap and has a breathy voice. This, in turn, has also a negative effect on intelligibility. However, the weighting factor for this feature in the regression formula (table VII) was rather low. Hence, it does not have the most important impact on intelligibility.

A similar reason holds for the relevance of the ratio of durations of the voiced sections in the recording and the entire recording. Highly irregular voices show a lower portion of harmonic segments and usually lower intelligibility.

The contribution of the standard deviation of jitter to automatic intelligibility rating is also based upon irregularities in voicing. It describes how much jitter is varying during the voiced sections of the text recording. A low value would point out a more or less regular voice while high values describe, in addition to varying $F_0$, that also the variation of $F_0$ is not regular. The standard deviation of $F_0$ over the entire sample incorporates information about irregularity in two ways: It measures the $F_0$ variation and takes into account the number of frames that were classified as voiced. When the standard deviation of $F_0$ was computed on the voiced frames only, the feature was not even selected into the best feature set.

### Measuring Intelligibility by Analyzing Text Recordings

For the purpose of 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 for the intelligibility of the patient. In a study on the German Post-Laryngectomy Telephone Test (PLTT, 30), however, it was shown that one naïve rater alone is not enough to achieve reliable results because the perceptually assessed results differ too much among the raters (31). Our intention is also to design an automatic support for speech therapy, so the reference data have to be obtained from trained listeners first.

There is another important reason for using a standard text: When automatic speech evaluation is performed for instance with respect to prosodic phenomena, such as word durations or percentage of voiced segments, then comparable results for all patients can only be achieved when all the patients read the same defined words or text. An inverse intelligibility test can no longer be performed then, and intelligibility has to be rated on a grading scale instead. However, the inter-rater correlation between an objective, automatic evaluation method and expert raters who rated intelligibility on a 5-point scale was well above 0.8 (11). For an objective, automatic version of the PLTT, correlations between the average naïve listener and the automatic results were in the same range (31). Hence, the text-based evaluation performed by trained listeners is as reliable as the inverse intelligibility test with naïve raters.

### Recognition System

For this study, polyphone-based recognizers were used because they yield better results in not severely deteriorated voices (11). 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 reduce the confounding factors on the recognition result to a minimum. The acoustic difference to real telephone speech was assumed marginal in comparison to the difference between normal and pathologic speech. The a-law companding function of the telephone channel was also shown to be not disadvantageous for human-machine correlation (11).

The WA and WR values were rather low. However, their absolute values are not crucial. It is their range that is important for an adequate representation of the perceptual ratings. The important measure for the success of the method is the human-machine correlation. The low WA and WR values were mainly caused by the training set of the speech recognizer consisting of the speech of young adults. On the one hand, there were not enough data to train a recognizer with pathologic speech only. Another aspect that has to be considered is the average age of the patients. There is an influence of age on automatic speech recognition results which was presented in a study by Wilpon and Jacobsen (32). When they trained their system with 19- to 34-year-old persons, which fits approximately the training speakers of our recognizers, they measured a word error rate of 2.4% on test speakers of the same age category. For test speakers of the age category 60+, the error rate was 4.7%. This was obtained in a digit recognition test, i.e. with a very small recognition vocabulary. For larger vocabularies, the error rates will be higher. However, the natural influence of age on the voice is in most cases much smaller than the influence that the patient's voice pathology has. For this reason, age is not the most critical factor in this method. Due to the limited amount of data, we could not examine age effects on pathologic speakers in detail so far. For the design of a recognizer for practical clinical application, this aspect will have to be taken into account. On the other hand, in view of the rehabilitation process of the patients, it is important to use a system that represents a reference for average normal speech (also with respect to age) in order to compute the difference between this reference and the pathology of the respective patient.

## Conclusion

Objective evaluation of intelligibility by automatic speech recognition and prosodic features is possible for close-talk speech and for telephone speech, even for pathologies with a broad range of degree. There is a high and significant correlation between subjective ratings and the automatic measures. The human-machine correlation is as high as the average inter-rater correlation among speech therapists. Hence, the method is suitable as basis for an automatic, objective, and easily applicable support for voice and speech assessment. In this article, it has been evaluated using speech data from laryngectomized patients. However, it can easily be transferred onto other types of pathology. Thus, it can serve as reliable "second opinion" for voice and speech therapy where usually one single therapist treats the patient.

**Declaration of Interest**

This work was funded by the German Cancer Aid (Deutsche Krebshilfe) under grant 107873. The responsibility for the content of this article lies with the authors. The authors report no declarations of interest.

**Acknowledgments**

We would like to thank Maria Schuster, Eva Uhl, Florian Hebel, and the speech therapists of the Department of Phoniatrics and Pediatric Audiology for obtaining the audio and perceptual evaluation data.

## References

1. Dejonckere PH, Bradley P, Clemente P, Cornut G, Crevier-Buchman L, Friedrich G, Van De Heyning P, Remacle M, Woisard V, Committee on Phoniatrics of the European Laryngological Society (ELS). A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. Guideline elaborated by the Committee on Phoniatrics of the European Laryngological Society (ELS). Eur Arch Otorhinolaryngol. 2001;258:77-82.
2. Dubuisson T, Dutoit T, Gosselin B, Remacle M. On the Use of the Correlation between Acoustic Descriptors for the Normal/Pathological Voices Discrimination. EURASIP Journal on Advances in Signal Processing; Analysis and Signal Processing of Oesophageal and Pathological Voices. 2009;2009:19 pages. DOI:10.1155/2009/173967.
3. Moran R, Reilly R, de Chazal P, Lacy P. Telephony-based voice pathology assessment using automated speech analysis. IEEE Trans Biomed Eng. 2006;53:468-77.
4. 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;47:160-8.
5. Vasilakis M, Stylianou Y. Voice pathology detection based eon [sic] short-term jitter estimations in running speech. Folia Phoniatr Logop. 2009;61:153-70.
6. Malyska N, Quatieri T, Sturim D. Automatic Dysphonia Recognition using Biologically-Inspired Amplitude-Modulation Features. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). Philadelphia, PA; 2005. vol. I, p.873-6.
7. Umapathy K, Krishnan S, Parsa V, Jamieson DG. Discrimination of pathological voices using a time-frequency approach. IEEE Trans Biomed Eng. 2005;52:421-30.
8. Halberstam B. Acoustic and Perceptual Parameters Relating to Connected Speech Are More Reliable Measures of Hoarseness than Parameters Relating to Sustained Vowels. ORL J Otorhinolaryngol Relat Spec. 2004;66:70-3.
9. Bäckström T, Lehto L, Alku P, Vilkman E. Automatic pre-segmentation of running speech improves the robustness of several acoustic voice measures. Logoped Phoniatr Vocol. 2003;28:101-8.
10. Hogikyan ND, Sethuraman G. Validation of an instrument to measure voice- related

quality of life (V-RQOL). J Voice. 1999;13:557-69.

11. Haderlein T, Riedhammer K, Nöth E, Toy H, Schuster M, Eysholdt U, Hornegger J, Rosanowski F. Application of Automatic Speech Recognition to Quantitative Assessment of Tracheoesophageal Speech in Different Signal Quality. Folia Phoniatr Logop. 2009;61:12-7.
12. Maier A, Haderlein T, Stelzle F, Nöth E, Nkenke E, Rosanowski F, Schützenberger A, Schuster M. Automatic Speech Recognition Systems for the Evaluation of Voice and Speech Disorders in Head and Neck Cancer. EURASIP Journal on Audio, Speech, and Music Processing. 2010;2010:7 pages.
13. 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;263:188-93. DOI: 10.1007/s00405-005-0974-6.
14. Kitzing P, Maier A, Åhlander VL. Automatic speech recognition (ASR) and its use as a tool for assessment or therapy of voice, speech, and language disorders. Logoped Phoniatr Vocol. 2009;34:91-6.
15. Ananthakrishnan S, Narayanan S. An Automatic Prosody Recognizer Using a Coupled Multi-Stream Acoustic Model and a Syntactic-Prosodic Language Model. In: Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP). Philadelphia, PA; 2005. vol. I, p.269-72.
16. Batliner A, Buckow J, Niemann H, Nöth E, Warnke V. The Prosody Module. In: Wahlster W, editor. Verbmobil: Foundations of Speech-to-Speech Translation. Berlin, Heidelberg, New York: Springer; 2000. p.106-21.
17. Chen K, Hasegawa-Johnson M, Cohen A, Borys S, Kim S-S, Cole J, Choi J-Y. Prosody dependent speech recognition on radio news corpus of American English. IEEE Trans Audio Speech Lang Process. 2006;14:232-45.
18. Gallwitz F, Niemann H, Nöth E, Warnke V. Integrated Recognition of Words and Prosodic Phrase Boundaries. Speech Commun. 2002;36:81-95.
19. Nöth E, Batliner A, Kießling A, Kompe R, Niemann H. Verbmobil: The Use of Prosody in the Linguistic Components of a Speech Understanding System. IEEE Trans Speech Audio Process. 2000;8:519-32.
20. Maier A. Speech of Children with Cleft Lip and Palate: Automatic Assessment. Dissertation, vol. 29 of Studien zur Mustererkennung. Berlin: Logos Verlag, 2009.
21. 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;264:1315-21.
22. World Medical Association. World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. 2008. http://www.wma.net/en/30publications/10policies/b3. Accessed 14 Jun 2011.
23. International Phonetic Association. Handbook of the International Phonetic Association. London: Cambridge University Press, 1999. DOI: 10.2277/0521637511.
24. Stemmer G. Modeling Variability in Speech Recognition. Dissertation, vol. 19 of Studien zur Mustererkennung. Berlin: Logos Verlag, 2005.
25. Arias-Londoño JD, Godino-Llorente JI, Markaki M, Stylianou Y. On combining information from modulation spectra and mel-frequency cepstral coefficients for automatic detection of pathological voices. Logoped Phoniatr Vocol. 2011;36:60-9.
26. Schukat-Talamazzini EG, Niemann H, Eckert W, Kuhn T, Rieck S. Automatic Speech Recognition without Phonemes. In: Proc. European Conf. on Speech Communication and Technology (Eurospeech). Berlin: European Speech Communication Association (ESCA), 1993. p.129-32.
27. Batliner A, Fischer K, Huber R, Spilker J, Nöth E. How to Find Trouble in Communication. Speech Commun. 2003;40:117-43.

28. Smola AJ, Schölkopf B. A tutorial on support vector regression. Statistics and Computing. 2004;14:199-222.
29. Witten I, Frank E, Trigg L, Hall M, Holmes G, Cunningham SJ. Weka: Practical machine learning tools and techniques with java implementations. In: Proc. ICONIP/ANZIIS/ANNES99 Future Directions for Intelligent Systems and Information Sciences. San Francisco: Morgan Kaufmann Publishers, 1999. p.192-6.
30. Zenner HP. The postlaryngectomy telephone intelligibility test (PLTT). In: Herrmann IF, editor. Speech Restoration via Voice Prosthesis. Berlin: Springer, 1986. p.148-52.
31. 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, editors. Proc. 10th Int. Conf. Text, Speech and Dialogue (TSD 2007). vol. 4629 of Lecture Notes in Artificial Intelligence. Berlin, Heidelberg, New York: Springer, 2007. p.238-45.
32. Wilpon JG, Jacobsen CN. A study of speech recognition for children and the elderly. In: IEEE Int. Conf. on Acoustics, Speech, and Signal Processing. Atlanta, Georgia, 1996. p.349-52.

## Tables

Table I
Perceptual intelligibility measures of 5 raters for 82 patients after partial laryngectomy (mean and standard deviation); "Avg." is the average over all raters.

| Rater | R1 | R2 | R3 | R4 | R5 | Avg. |
|---|---|---|---|---|---|---|
| Headset | 2.98SD1.22 | 2.73SD1.30 | 3.00SD0.99 | 2.91SD1.28 | 2.91SD1.15 | 2.91SD1.07 |
| Telephone | 3.21SD1.20 | 3.13SD1.39 | 3.38SD1.01 | 3.23SD1.26 | 3.32SD1.29 | 3.25SD1.10 |

Table II
Word accuracy (WA) and word correctness (WR) values for 82 patients after partial laryngectomy

| Data | Sample Freq. | Measure | Mean | St. Dev. | Minimum | Maximum |
|---|---|---|---|---|---|---|
| Headset | 16 kHz | WA | 51.8 | 19.2 | 0.0 | 82.4 |
| | | WR | 57.2 | 16.5 | 14.9 | 84.3 |
| Telephone | 8 kHz | WA | 47.0 | 19.6 | -2.7 | 79.6 |
| | | WR | 53.9 | 17.0 | 8.6 | 83.3 |

Table III
Inter-rater correlation between a single raters' intelligibility scores and the average of the 4 other raters, evaluated on 82 recordings; "Avg." is the average over all raters.

| Rater | R1 | R2 | R3 | R4 | R5 | Avg. |
|---|---|---|---|---|---|---|
| r(Headset) | 0.79 | 0.85 | 0.83 | 0.84 | 0.88 | 0.84 |
| r(Telephone) | 0.85 | 0.82 | 0.87 | 0.79 | 0.85 | 0.84 |

Table IV
Intra-rater correlation for the raters' intelligibility scores, evaluated on 44 recordings; "Avg." is the average over all raters.

| Rater | R1 | R2 | R3 | R4 | R5 | Avg. |
|---|---|---|---|---|---|---|
| r(Headset) | 0.73 | 0.86 | 0.86 | 0.85 | 0.85 | 0.83 |
| r(Telephone) | 0.90 | 0.78 | 0.83 | 0.82 | 0.83 | 0.83 |

Table V
Intra-rater correlation between the headset and telephone recordings for the raters' intelligibility scores, evaluated on 82 recordings; "Avg." is the average over all raters.

| Rater | R1 | R2 | R3 | R4 | R5 | Avg. |
|---|---|---|---|---|---|---|
| r(Headset vs. Telephone) | 0.71 | 0.77 | 0.81 | 0.82 | 0.75 | 0.77 |

Table VI
Human-machine correlation for headset and telephone recordings; the left column specifies the set of measures that was used to compute the correlation to the average human listener.

| Measures | Headset | Telephone |
|---|---|---|
| WA only | -0.57 | -0.69 |
| WR only | -0.62 | -0.75 |
| Best set | +0.79 | +0.86 |

Table VII
Weighting factors of the elements of the best feature set for human-machine correlation; the higher the absolute weighting factor is in the regression formula between perceptual and automatic evaluation, the more important is the feature.

| Feature | Headset | Telephone |
|---|---|---|
| Duration of silent pause before word | 0.291 | 0.221 |
| Standard deviation of $F_0$ (all frames) | 0.616 | 0.350 |
| Standard deviation of jitter | 0.243 | 0.332 |
| Ratio of durations: voiced sections/entire recording | -0.916 | -0.775 |
| WR | -0.476 | -0.641 |