Language-Independent Automatic Evaluation of Intelligibility of Chronically Hoarse Persons

Tino Haderlein[1/2], Catherine Middag[3], Jean-Pierre Martens[3], Michael Döllinger[1/4], Elmar Nöth[2/5]

[1]Universitätsklinikum Erlangen
Phoniatrische und pädaudiologische Abteilung
Bohlenplatz 21
91054 Erlangen
Germany

[2]Lehrstuhl für Mustererkennung
Universität Erlangen-Nürnberg
Martensstraße 3
91058 Erlangen
Germany

[3]Vakgroep voor Elektronica en Informatiesystemen (ELIS)
Universiteit Gent
Sint-Pietersnieuwstraat 41
9000 Gent
Belgium

[4]Communication Sciences and Disorders Department
Louisiana State University (LSU)
63 Hatcher Hall
Baton Rouge, LA 70803
USA

[5]Electrical & Computer Engineering Department
Faculty of Engineering
King Abdulaziz University
Jeddah 21589
Saudi Arabia

Short Title: Language-independent Automatic Evaluation of Intelligibility

Corresponding author:
Tino Haderlein
Universitätsklinikum Erlangen
Phoniatrische und pädaudiologische Abteilung
Bohlenplatz 21
91054 Erlangen
Germany
E-Mail: Tino.Haderlein@cs.fau.de
Phone: +49 9131 85 27872
Fax: +49 9131 303811

Abstract:
Objective: Automatic intelligibility assessment using automatic speech recognition is usually language-specific. In this study, a language-independent approach is proposed. It uses models that are trained with Flemish speech, and it is applied to assess chronically hoarse German speakers. The research questions here are: is it possible to construct suitable acoustic features that generalize to other languages and a speech disorder, and is the generated model for intelligibility also suitable for specific subtypes of that disorder, i.e. functional and organic dysphonia?
Patients and Methods: 73 German speaking persons with chronic hoarseness read the text “Der Nordwind und die Sonne”. Perceptual intelligibility scores were used as ground truth during the training of an automatic model that converts speaker-level acoustic measurements into intelligibility scores. Cross-validation is used to assess model performance.
Results: The inter-rater agreement for all patients (n=73) and for the functional and organic dysphonia subgroups (n=45 and n=24) are r=0.82, r=0.83, and r=0.75, respectively. The automatic assessment based on phonologically-based acoustic models revealed correlations between perceptual and automatic intelligibility ratings of r=0.79 (all patients), r=0.78 (functional dysphonia), and r=0.80 (organic dysphonia).
Conclusion: Automatic, objective measurement of intelligibility is a valuable instrument in an evidence-based clinical practice.

Keywords: Acoustic analysis - Intelligibility - Perceptual rating - Running speech - Voice disorders - Chronic hoarseness - Objective analysis - Phonologic features

1. Introduction

Subjective-perceptual voice and speech evaluation cannot fulfill the requirements of evidence-based medicine [1]. For instance, it is problematic with respect to differences in degrees of experience among the examiners [2], because every person has used previously processed data to develop an individual way of judging new data. For this reason, subjective evaluation evolves continuously, and it therefore has to be replaced or at least supplemented by objective, automated methods. However, until recently, the latter were usually restricted to voice quality measurements on sustained vowels or single phones, i.e. speech stimuli that differ considerably from natural speech encountered in realistic communication settings [3, 4, 5, 6]. Important speech criteria, such as intelligibility, cannot be obtained in this way and require much more elaborate solutions.

Intelligibility has been identified as one of the most important aspects of voice and speech assessment [7, 8, 9, 10]. In clinical practice, it is usually evaluated perceptually, but automatic intelligibility assessment tools employing an automatic speech recognition (ASR) system have recently emerged and have shown big potential [11, 12]. However, an ASR system encompasses acoustic models for the basic speech sounds (e.g. phonemes). These models are usually trained on non-distorted speech from one particular language. During speech assessment, a test subject reads a particular text, and the acoustic models establish how well the tested speech compares to the speech that would have been obtained from a “normal” person reading the same text (the “expected” speech). Consequently, the system can only be used to assess speech of the language that was used for acoustic model training, and it may be affected by reading errors made by the tested subject since such errors also affect the “expected speech”.

In order to further extend automatic speech assessment to spontaneous speech assessment, where no transcriptions of the speech are available anymore, Bocklet et al. [13] proposed to adopt a speaker verification approach. In that approach, the statistical distribution of the spectra of 20 milliseconds speech frames is represented by a Gaussian

Mixture Model (GMM), and the parameters of that GMM constitute the speaker features, i.e. measures that are characteristic for one specific person. Intelligibility is then inferred from the discrepancies between these speaker features and the corresponding features of a normal speaker set. If a GMM is trained on a sufficiently long speech sample, it is deemed to be largely text-independent. Middag et al. [14] propose an alternative approach. They train a neural network to convert a speech spectrum into a vector of phonological scores, and they subject these scores to a holistic analysis. Each score represents the posterior probability of one phonological class (e.g. central vowel), i.e. the probability that the currently processed section of the speech sample represents a phone of the respective class. By examining which fraction of the time it is above a certain threshold, how long on average are the intervals in which it stays above or below that threshold, etc., one obtains a number of speaker features that can be compared to the corresponding features derived from normal speech. These speaker features are regarded text-independent for the most part.

In the present study, we investigate the latter approach [14] because we already demonstrated that a phonological feature extractor trained on Flemish speech constitutes a suitable basis for assessing the intelligibility of German speaking patients with various voice pathologies [15]. However, the inputs of the phonological feature extractor [16] are Mel-Frequency Cepstrum Coefficients (MFCCs) that just represent the spectral envelope. Therefore, in the present study we supplement the phonological speaker features with the AMPEX features proposed by Moerman et al. [10]. They originate from a holistic analysis of the frame-level volume, fundamental frequency, and voicing evidences. Note that this analysis can be conducted on arbitrary speech, irrespective of the language that is spoken, and that it therefore fits well in the multilingual setting of the present study.

The study is part of a research project on the analysis of chronic hoarseness. Elderly persons were reported to have 29% of a point prevalence and a lifetime incidence of 47% for a voice disorder. Hoarseness is one of the most important symptoms, appearing in 78% of the cases [17]. For this reason, a specific model for the intelligibility of hoarse persons was supposed to be created. Chronic hoarseness can have different causes, such as functional problems, organic variations, or laryngitis. Therefore, we tested whether one single model for all types of hoarseness is sufficient to evaluate the most prevalent subtypes reliably.

The focus of this work is on the importance of the different phonological and prosodic dimensions for the prediction of intelligibility ("prediction" means a value computed by a statistical model in this case, not a prognosis on events in the future). More in particular, the following questions will be addressed:

- Do phonological feature extractors trained on undistorted Flemish speech enable a reliable assessment of the intelligibility deficiencies observed in a group of German persons with chronic hoarseness?
- Which measures (features) computed from which phonological dimensions are most important for creating a good intelligibility model?
- Do the obtained features work equally well for different speech pathology subgroups?

2. Material and Methods

2.1. Subjects and Test Samples

73 German persons with chronic hoarseness (24 men and 49 women) between 19 and 85 years of age participated in this study (Table 1). The average age was 48.3, the standard deviation was 16.8 years. Patients suffering from cancer, patients with hearing problems, and persons who were not able to read a text from a sheet of paper were excluded. No other pre-selection was made. The most common pathologies were grouped into functional (n=45)

and organic dysphonia (n=24; Table 2). The four remaining persons suffered from laryngitis and were not evaluated separately. Each person read the text “Der Nordwind und die Sonne” (“The North Wind and the Sun”), a phonetically rich text with 108 words (71 distinct) and 172 syllables. This text is frequently used in medical speech evaluation in German-speaking countries and is acknowledged as a standard text for these purposes by the International Phonetic Association [18]. The text was read as one passage, recorded with a headset microphone AKG C 420 (AKG Acoustics, Vienna, Austria), sampled at a frequency of 16 kHz and digitally stored in 16 bit pulse code modulation (PCM).

The study respected the principles of the World Medical Association (WMA) Declaration of Helsinki on ethical principles for medical research involving human subjects and has been approved by the ethics committee of the university clinics in Erlangen. All patients gave their informed consent to participate in this study.

2.2. Subjective Evaluation

A group of five voice professionals was instructed to estimate the intelligibility of the patients while listening to play-backs of the recordings. The samples were presented in random order. 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. An averaged mark, expressed as a floating point value, was calculated for each patient as the mean of the single scores. These marks served as ground truth in our experiments.

2.3. Objective Evaluation

The objective evaluation of intelligibility is a four-step procedure. The pre-processing stage produces a spectro-temporal representation of the acoustic signal. The phonological analysis stage converts the spectral envelope features into phonological features that expose the phonological properties of the subsequent speech frames. The speaker feature extraction stage then performs a holistic analysis of the phonological features and creates a compact set of speaker features. In the intelligibility assessment stage, the speaker features are converted into an intelligibility score by means of a regression model that is referred to as an intelligibility model.

2.3.1. Pre-processing

During the pre-processing stage, a stream of Mel-frequency cepstral coefficients (MFCCs) is extracted from the recording. MFCCs are standard features in automatic speech processing [16], because they describe the spectro-temporal evolution of speech in a compact way. Every 10 milliseconds, a speech frame of 25 milliseconds centered around the current timestamp is Hamming-windowed and analyzed. The analysis returns 12 MFCCs and an energy value. The MFCCs describe the shape of the spectrum in decibels as a function of the logarithm of the frequency. The energy value is computed from the amplitudes in the time domain.

In parallel, the recordings are also analyzed by means of the auditory model proposed by Van Immerseel & Martens [19]. Every 10 milliseconds, this analysis generates a voicing evidence describing the degree of regularity in the voiced frames, a voiced-unvoiced decision, and a fundamental frequency value. Even if the voicing decision is “unvoiced”, the analysis still returns the “most likely” fundamental frequency. These features are called prosodic features.

2.3.2. Phonological Analysis

Per 10 milliseconds, a window of five MFCC vectors (representing the considered frame as well as the two preceding and the two succeeding frames) is converted into a vector of phonological scores. Given the MFCC inputs, each score represents the probability that the considered frame belongs to a sound that belongs to a certain phonological class. Each phonological class refers to a group of basic sounds sharing a certain property. In the present study, we consider the 14 properties that were also used in [14, 20], because they do not require analysis of longer speech segments. There is one property of the vocal source (voicing), four properties referring to manner of articulation (silence, consonant-nasality, vowel-nasality, turbulence), six properties referring to place of consonant articulation (labial, labio-dental, alveolar, velar, glottal, palatal), and three properties referring to vowel features (height, place, lip rounding). Vowel height is ordered from low to high, vowel place is ordered from front to back. Most of them are of a binary nature, but a few (vowel height and place) are of a ternary nature.

The envisaged conversion is achieved by means of a feed-forward neural network, as described in [14, 20]. The network is trained on a corpus of Flemish continuous speech utterances elicited from speakers with no apparent voice disorders. Each utterance was automatically segmented and labeled into basic sounds and silences, and these labels were converted to target values for the phonological properties. In the case of a binary property, this target value is 0 or 1. In the case of a ternary variable, the target value can be 0, 0.5, or 1. During training a frame belonging to a consonant will not contribute to the training of the network outputs referring to a vowel feature and vice versa.

2.3.3. Speaker Feature Extraction

The frame-level phonological and prosodic features are each subjected to a holistic analysis which does not need any precise knowledge of the text that was read.

Prosodic Features (AMPEX)
The frame-level prosodic features are converted into 8 so-called AMPEX features, as described in [10]. The voicing evidence and the signal loudness (see [19, 21]) are used to label the frames as voiced/unvoiced and as speech/silence, and to locate pauses, defined as intervals of more than 200 ms long. Based on these classifications, the AMPEX feature extractor computes the features listed in Table 3. They can be grouped into voicing-related parameters (e.g. the percentage of speech frames classified as voiced) and F_0-related features (e.g. average jitter of the fundamental frequency F_0 in voiced frames). The features were computed for the whole length of each speech sample. In earlier studies it has been shown that supplementing phonological features with these F_0-and-voicing related speaker characteristics leads to enhanced intelligibility prediction [20].

Phonological Features (ALF-PLFs)
The frame-level values of one phonological feature (e.g. the vowel-nasality feature) form the samples of a temporal pattern, and the speaker feature extraction performs an analysis of the static and dynamic properties of this pattern, as described in [14]. Each pattern gives rise to 18 properties, and all samples of a pattern have a value between 0 and 1. Among the static properties are the mean and standard deviation of the sample distribution, the percentage and mean values of samples with a value between 0 and 0.33 (low), between 0.33 and 0.66 (intermediate), and between 0.66 and 1 (high). Among the dynamic features are the mean value of the peaks and valleys (maxima, minima) in the pattern, the mean lengths of the sections with the aforementioned value intervals, and the mean time to reach a maximum or minimum value. There are 14 phonological patterns, each of them specified by its relevance and presence, and 18 speaker features extracted for each of these patterns. Hence, 504 speaker features are computed for each speech sample. Note that the holistic

analysis of the vowel features is restricted to the time intervals where the vowel evidence is larger than 0.5. For the consonant features, this holds accordingly.

Since these phonological features do not require a transliteration of the spoken text, they will be denoted as alignment-free phonological features (ALF-PLFs), in accordance with [20].

2.3.4. Intelligibility Model – Experimental Setup

Once every speaker is characterized by a set of speaker features, a regression model is trained. The aim is to minimize the mean-squared difference across training speakers between the model output, i.e. the computed intelligibility values, and the target output, i.e. the average perceptual mark.

In order to avoid over-fitting to the training data, a 5-fold cross-validation setup was used, in which always four-fifths of the utterances form a training partition meant for model training. The remaining utterances form a test partition meant for model testing.

The models were trained using support vector regression (SVR, [22]). The underlying Support Vector Machine (SVM) employs a Gaussian kernel. During training on a particular training fold, the learning parameters (kernel parameters, fault threshold) are retrieved from a grid search which is based on an internal 5-fold cross validation scheme. In such a scheme, the aforementioned training partition for the models is split up again. Models are trained on four-fifths of the training partition and evaluated on the remaining fifth of that partition. The experiment is repeated five times for five subdivisions of the training partition. Once the learning parameters are fixed, a new SVM is trained on the full training partition and evaluated on the corresponding test partition.

In order to determine which features are most relevant, we ranked the features according to their absolute weights in the SVR models. To that end we divide the raw feature weights emerging from one training by the maximum over all feature weights, and we take the means of these normalized feature weights over the five folds. The theoretical maximum of this mean is then equal to 5.

2.3.5. Agreement within the Rater Group and between Raters and Automatic Evaluation

Pearson's correlation coefficient r was used to measure the correlation between the computed intelligibility values and the human ratings. For the agreement within the rater group, Pearson's r and Krippendorff's α [23] were determined. Per definition, Krippendorff's α can be computed for the whole group all at once. In order to achieve a corresponding value for r, the correlation between one rater's scores and the average scores of the remaining four raters were computed. All five of these "four vs. one" scores were then averaged to form the final correlation value. For the human-machine agreement, Krippendorff's α was not used since it would have required a mapping of the real-valued regression results to integer numbers, which would have introduced another source of error.

3. Results

3.1. Subjective Evaluation

The perceptual scores for the 73 patients covered the whole range of the 5-point scale (Table 4), with an average mark of 2.51 for the whole group. Intelligibility of the persons with functional dysphonia was regarded better (average: 2.27) than that of the persons with organic dysphonia (3.06). The distribution of the judgments is depicted in Figure 1. Fifteen persons with functional dysphonia were rated to have a very high intelligibility of below 1.5

on the average. For the organic dysphonia, this was not a single person. The inter-rater correlations for the entire patient group, the group with functional dysphonia, and the organic dysphonia group were r=0.83, r=0.82, and r=0.75, respectively (p<0.001). The corresponding values for Krippendorff's α were 0.70, 0.67, and 0.64, respectively.

3.2. Objective Evaluation

The human-machine correlation between the average perceptual intelligibility ratings and the automatically computed values was r=0.79 for the whole patient group. For the functional dysphonia group, it was r=0.78, for the organic dysphonia group, r=0.80 was achieved (p<0.001). The figures confer well with the human inter-rater agreement.

Table 5 shows the 10 features which had the highest weight in the models and which can thus be considered most important for predicting the intelligibility of the considered type of dysphonic speakers. Among the 10 features, we find four AMPEX features (AVE, PVS, PVF and PVFU) and six ALF-PLF features. Two of the latter features are related to turbulence, two to vowel-consonant distinction, one to consonant nasality and one to the presence of silences in the speech.

4. Discussion

The main target of this study is reaching good predictions of intelligibility. In this respect, our results show that the model predictions for the entire patient group correlate almost as much with the ground truth (r=0.79) as the scores of an arbitrary human rater (r=0.83) who actually contributed to that ground truth. Moreover, the human-machine correlation is very similar when measured on the entire speaker group, the functional dysphonia subgroup, and the organic dysphonia subgroup. The two subgroups cover about the same age range (Table 1), and they were recorded with the same microphone at the same place. The speakers all live in the same region and do not suffer from other impairment than hoarseness. It is not possible to form two groups which are identical in every single aspect, but according to our experience, we suppose that a large portion of the differences between their evaluation results emerges from different causes of hoarseness. The human-machine correlations, on the other hand, can also be influenced by the different size of the subgroups. Nevertheless, the results suggest that our automatic intelligibility assessment can deal with different types of hoarseness. To find more substantial evidence for this, we could have compared our general intelligibility model with specific models trained on one speaker subset. However, due to the small size of the subgroups, we would not have been able to construct specific models of the size of the general model, and the conclusions of the experiment would have been debatable.

Support Vector Regression on all 504 ALF-PLF and 8 AMPEX features revealed the 10 most relevant features for modeling the human reference. Four of them describe the speaker's control over his or her phonation: voicing evidence in voiced frames, the percentage of frames classified as voiced, the percentage of speech frames classified as voiced, and the percentage of frames classified as voiced with an unreliable F_0. Among the phonological features, the most important ones are turbulence-related. This was expected since dysphonic speakers are known to produce turbulence in vowels that are normally supposed to exhibit a harmonic spectrum. The selection of the silence property may stem from a higher speaking effort for dysphonic speakers which leads to more breathing noise and thus to "filled" rather than clear silences in the speech. This has been shown similarly in the effort ratings for laryngectomized persons [11]. The importance of the feature concerning the relevance of the nasality of consonants may have a similar reason.

Earlier studies using speech recognition techniques have also shown that the durations of pauses and voiced segments within words are highly correlated with intelligibility ratings, at least for the case of partially and totally laryngectomized persons [11, 12]. The results of our current study confirm these findings. In future experiments, the selection procedure for the most relevant features will be enhanced based upon feature ranks according to the amount of variance in the computed intelligibility score across the training samples they represent. According to our experience and the mentioned former studies, however, we do not expect significant changes in the results.

The procedure of perceptive evaluation, which was used for this study, may raise the question whether the raters really evaluated intelligibility. The way of evaluation was supposed to depict the methods that are usually applied in therapy sessions. The raters were clearly instructed to evaluate intelligibility instead of voice quality, because it is known that the degree of voice distortion influences the rating of intelligibility [24]. It is very difficult, however, to exclude this effect in clinical practice where intelligibility is often not evaluated as a percentage of correctly understood words, because these exact tests are time-consuming. Additionally, a percentage scale is too detailed to be relevant for therapy suggestions. The percentage values would very likely be grouped into a small number of intervals with a certain decision for therapy for each of them. For this reason, we decided to instruct the therapists to rate intelligibility in five classes right from the beginning. It is obvious that these labeled classes may not be assigned uniformly by the raters due to certain bottom or floor effects, which actually makes the conversion to integer numbers a non-linear operation. However, in the same way we regard it as very likely that the effect on communication success by differences in percentage intelligibility is also not equally distributed. A comprehensive study on these effects is not the topic of this work, but we believe that the difference between 30% and 40% of understood words, for instance, will cause another degree of information loss than between 90% and 100%. The average value of the ratings of several raters was used in order to get a representative evaluation, not a single one with personal bias. Some researchers prefer the consensus method, where the raters agree on a common rating. But this does not reflect the average of independent ratings, since some of the involved persons may neglect their own impression and rather choose a label which is more consistent with the others.

Another point of criticism may be the use of a standard text for this study. This was done in order to be consistent with former studies, where ASR-based evaluation methods were used [11, 12, 13]. They required the same words spoken by every patient, and the reference evaluation was obtained in the same way. On the other hand, this ensures the same vocabulary and number of words in every speech sample to the most possible extent. Hence, human perception is not influenced by variations in those aspects.

It has been shown in this study that the proposed phonological approach is essentially language-independent: phonological models trained on normal Flemish speech can assess German pathologic speech. Experiments with more dissimilar language pairs should be conducted in order to further substantiate the above statement. Also within one language, the features may be partly determined by the underlying phoneme and word frequencies. Consequently, variations in the read text, e.g. variations due to reading errors and hesitations, do increase the inter-subject variability. Preliminary experiments with the data presented in this study have shown that this influence is on average not significant, but it should nevertheless be examined more in detail. A remaining challenge for future work is to develop a methodology that is capable of predicting intelligible of speech spoken over different media: plain old telephones, mobiles and internet (e.g. Skype) [12, 25]. This is important since after all, these media presently constitute important communication means.

5. Conclusion

Until now, there is no generally accepted objective method for the evaluation of speech intelligibility. Here, we present an automatic evaluation system that can assess speech spoken in languages that were not used during the development of the system. It is achieved by the analysis of running speech which corresponds to the everyday use of voices. Furthermore, the test speaker set consisted of hoarse persons, and the training speakers did not show any voice pathology. The generated model for intelligibility can even be applied to subgroups of chronic hoarseness without model adaptation. The correlation between the automatic analysis and a human reference is on the same level as the inter-rater correlation within a group of experienced voice pathologists. The additional advantage of the system is that its results are reproducible, since it is based on acoustic measures. It cannot be fully objective, since intelligibility is a subjective measure. However, the reference for the training of the computed model was a representative average opinion of a panel of voice and speech therapists. Hence, such a model can serve as the basis for a future "objective", evidence-based approach for diagnostics and therapy support.

Acknowledgments
This work was partially funded by the Else Kröner-Fresenius-Stiftung, Bad Homburg v.d.H., Germany, under grant 2011_A167. It was also realized with support of "Kom op tegen Kanker", the campaign of the Vlaamse Liga tegen Kanker VZW and of the The Netherlands Cancer Institute / Antoni van Leeuwenhoek Hospital (Amsterdam).

Conflict of Interest
The authors state that there is no conflict of interest.

References

[1] Biddle AK, Watson LR, Hooper CR, Lohr KN, Sutton SF: Criteria for Determining Disability in Speech-Language Disorders. Summary, Evidence Report/Technology Assessment: Number 52. AHRQ Publication No. 02-E009. Rockville, MD, Agency for Healthcare Research and Quality, 2002.

[2] Eadie TL, Kapsner M, Rosenzweig J, Waugh P, Hillel A, Merati A: The Role of Experience on Judgments of Dysphonia. J Voice 2010;24:564-573.

[3] de Bruijn MJ, ten Bosch L, Kuik DJ, Quené H, Langendijk JA, Leemans CR, Verdonck-de Leeuw IM: Objective acoustic-phonetic speech analysis in patients treated for oral or oropharyngeal cancer. Folia Phoniatr Logop 2009;61:180-187.

[4] Fraile R, Sáenz-Lechón N, Godino-Llorente JI, Osma-Ruiz V, Fredouille C: Automatic Detection of Laryngeal Pathologies in Records of Sustained Vowels by Means of Mel-Frequency Cepstral Coefficient Parameters and Differentiation of Patients by Sex. Folia Phoniatr Logop 2009;61:146-152.

[5] van Gogh C, Festen J, Verdonck-de Leeuw I, Parker A, Traissac L, Cheesman A, Mahieu H: Acoustical analysis of tracheoesophageal voice. Speech Communication 2005;47:160-168.

[6] 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.

[7] Ainsworth W, Singh W. Perceptual Comparison of Neoglottal, Oesophageal and Normal Speech. Folia Phoniatr 1992;44:297-307.

[8] van As CJ, Koopmans-van Beinum FJ, Pols LC, Hilgers FJ: Perceptual evaluation of tracheoesophageal speech by naive and experienced judges through the use of semantic differential scales. J Speech Lang Hear Res 2003;46:947-959.

[9] Bellandese M, Lerman J, Gilbert H: An Acoustic Analysis of Excellent Female Esophageal, Tracheoesophageal, and Laryngeal Speakers. J Speech Lang Hear Res 2001;44:1315-1320.

[10] 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.
[11] Haderlein T. Automatic Evaluation of Tracheoesophageal Substitute Voices. Berlin, Logos Verlag, 2007.
[12] Haderlein T, Nöth E, Batliner A, Eysholdt U, Rosanowski F. Automatic intelligibility assessment of pathologic speech over the telephone. Logoped Phoniatr Vocol 2011;36:175-181.
[13] Bocklet T, Riedhammer K, Nöth E, Eysholdt U, Haderlein T: Automatic intelligibility assessment of speakers after laryngeal cancer by means of acoustic modeling. J Voice 2012; 26:390-397.
[14] Middag C, Saeys Y, Martens J-P: Towards an ASR-free objective analysis of pathological speech; in Proc. Interspeech. International Speech Communication Association, 2010, pp 294-297.
[15] Middag C, Bocklet T, Martens J-P, Nöth E: Combining phonological and acoustic ASR-free features for pathological speech intelligibility assessment; in Proc. Interspeech. International Speech Communication Association, 2011, pp 3005-3008.
[16] Davis SB, Mermelstein P: Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences. IEEE Trans Acoust Speech Signal Process 1980;28:357-366.
[17] Roy N, Stemple J, Merrill RM, Thomas L. Epidemiology of voice disorders in the elderly: preliminary findings. Laryngoscope 2007;17:628-633.
[18] International Phonetic Association: Handbook of the International Phonetic Association. Cambridge, Cambridge University Press, 1999.
[19] van Immerseel L and Martens J-P: AMPEX Disordered Voice Analyzer [computer program]. Digital Speech and Signal Processing research group, Ghent University, Ghent, Belgium. Available: http://dssp.elis.ugent.be/downloads-software. Last visited May 14, 2014.
[20] Middag C. Automatic Analysis of Pathological Speech. PhD thesis. Ghent University, Ghent, Belgium, 2012.
[21] Van Immerseel LM, Martens JP. Pitch and voiced/unvoiced determination with an auditory model. J Acoust Soc Am 1992;91:3511-3526.
[22] Smola AJ, Schölkopf B: A tutorial on support vector regression. Statistics and Computing 2004;14:199-222.
[23] Krippendorff K. Content analysis: An introduction to its methodology, 3rd ed. Thousand Oaks, CA, Sage, 2013.
[24] Weismer G, Martin R. Acoustic and perceptual approaches to the study of intelligibility.; in Kent R (ed.): Intelligibility in Speech Disorders. Philadelphia, John Benjamins Publishing Co, 1992. pp 67–118.
[25] Lin E, Hornibrook J, Ormond T: Evaluating iPhone recordings for acoustic voice assessment. Folia Phoniatr Logop 2012;64:122-130.

Tables

Table 1. Age statistics for the automatically evaluated patient groups

	n	men	women	avg. age	st.dev. (age)	min. age	max. age
total group	73	24	49	48.3	16.8	19	85
functional	45	13	32	47.1	16.3	20	85
organic	24	9	15	52.2	15.6	25	79

Table 2. Diagnoses within the speaker groups

subgroup	diagnosis	n
functional dysphonia	hyperfunctional dysphonia	23
	hypofunctional dysphonia	8
	combined functional dysphonia	14
organic dysphonia	organic dysphonia	9
	organic dysphonia + paresis	1
	spasmodic dysphonia	1
	vocal fold polyp	6
	paresis	4
	paresis + Reinke's edema	1
	Reinke's edema	2
laryngitis	laryngitis	2
	laryngitis + functional dysphonia	1
	laryngitis + organic dysphonia	1

Table 3. The AMPEX features (for details, see [21])

feature	description
PVF	percentage of **all** frames in the recording that were labeled voiced
PVS	percentage of **speech** frames that were labeled voiced
AVE	average voicing evidence in **voiced** frames
PVFU	percentage of **voiced** frames with an unreliable F_0
Jit	average F_0-jitter in **voiced** frames
Jc	average F_0-jitter in **voiced** frames with a reliable F_0
VL90	90th percentile (in seconds) of the voiced fragment durations
Tmax	duration (in seconds) of the longest **speech** fragment (not interrupted by a pause)

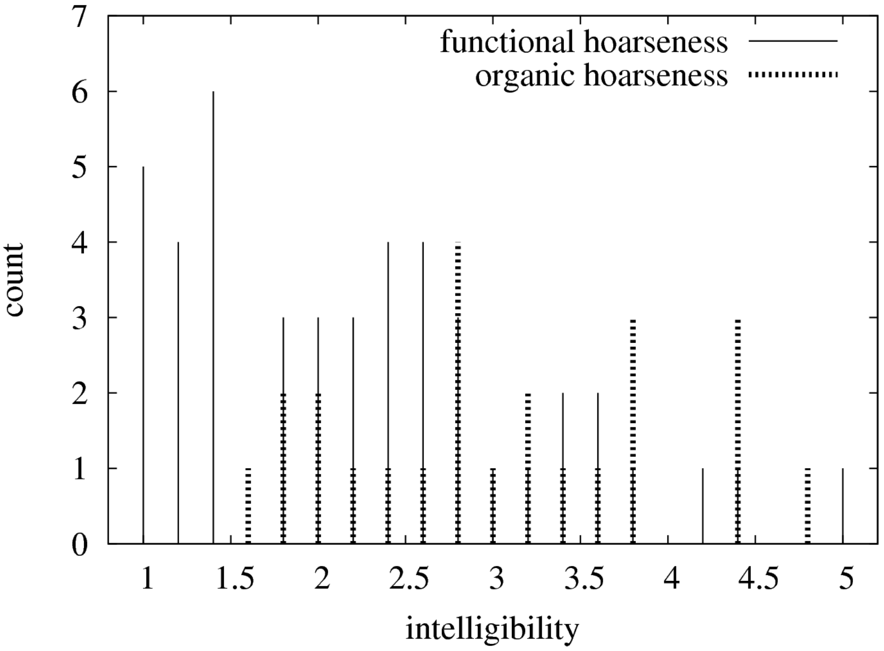
Table 4. Perceptual evaluation results (intelligibility on a 5-point scale)

	avg.	st.dev.	min.	max.	inter-rater correlation r	Krippendorff's α
total group	2.51	1.02	1.00	5.00	0.82	0.70
functional	2.27	1.00	1.00	5.00	0.83	0.67
organic	3.06	0.91	1.60	4.80	0.75	0.64

Table 5. Weights of the ten most important features in the intelligibility model

feature group	feature	weight
ALF-PLF	average maximum of turbulence being present	4.58
AMPEX	AVE – average voicing evidence in voiced frames	4.45
AMPEX	PVS – percentage of speech frames that were labeled voiced	4.27
ALF-PLF	average minimum of relevance of nasality of consonants	3.67
ALF-PLF	average minimum of absence of turbulence	3.63
ALF-PLF	average evidence for absence of vowels	3.63
AMPEX	PVF – percentage of all frames in the recording that were labeled voiced	3.52
AMPEX	PVFU – percentage of voiced frames with an unreliable F_0	3.27
ALF-PLF	average evidence for vowel dimension low	3.05
ALF-PLF	average evidence for the absence of silence	2.98

Figure 1. Perceptual evaluation of intelligibility for the functional (n=45) and the organic dysphonia group (n=24); the y-axis denotes the number of speakers with the respective intelligibility rating.


functional hoarseness
organic hoarseness
count
intelligibility