Automatic Intelligibility Assessment of Speakers After Laryngeal Cancer by Means of Acoustic Modeling

*,†Tobias Bocklet, †Korbinian Riedhammer, †Elmar Nöth, *Ulrich Eysholdt, and *,†Tino Haderlein, *†Erlangen, Germany

Summary: Objective. One aspect of voice and speech evaluation after laryngeal cancer is acoustic analysis. Perceptual evaluation by expert raters is a standard in the clinical environment for global criteria such as overall quality or intelligibility. So far, automatic approaches evaluate acoustic properties of pathologic voices based on voiced/unvoiced distinction and fundamental frequency analysis of sustained vowels. Because of the high amount of noisy components and the increasing aperiodicity of highly pathologic voices, a fully automatic analysis of fundamental frequency is difficult. We introduce a purely data-driven system for the acoustic analysis of pathologic voices based on recordings of a standard text.

Methods. Short-time segments of the speech signal are analyzed in the spectral domain, and speaker models based on this information are built. These speaker models act as a clustered representation of the acoustic properties of a person's voice and are thus characteristic for speakers with different kinds and degrees of pathologic conditions. The system is evaluated on two different data sets with speakers reading standardized texts. One data set contains 77 speakers after laryngeal cancer treated with partial removal of the larynx. The other data set contains 54 totally laryngectomized patients, equipped with a Provox shunt valve. Each speaker was rated by five expert listeners regarding three different criteria: strain, voice quality, and speech intelligibility.

Results/Conclusion. We show correlations for each data set with r and $\rho \geq 0.8$ between the automatic system and the mean value of the five raters. The interrater correlation of one rater to the mean value of the remaining raters is in the same range. We thus assume that for selected evaluation criteria, the system can serve as a validated objective support for acoustic voice and speech analysis.

Key Words: Laryngeal cancer–Total laryngectomy–Provox shunt valve–Voice quality–Intelligibility–Perceptual evaluation–Acoustic analysis.

INTRODUCTION

Laryngeal cancer affects the naturalness of a person's voice and has a major impact on the communication skills of affected persons.[1] In the United States, 34% of all workers have voice-dependent occupations.[2] In urban areas, the percentage is more than 87.5%. The prevalence of communication disorders in general is 5–10%. Persons affected with severe speech disabilities are more often found to be unemployed or in a lower economic class than people with hearing loss or other disabilities. Thus, voice and communication disorders have a major impact on the economy.[3] The economic effect is amplified by the influence on the patients' quality of life,[4–7] social, and psychosocial effects of voice and speech disorders.[8,9] Therefore, the rehabilitation of patients with communication disorders is of high clinical and economical interest.

To improve the voice quality and achieve intelligible and acceptable speech, reconstructive surgery followed by therapy is performed. The evaluation of voice and articulation capabilities after laryngeal rehabilitation has to be based on subjective and objective methods.[10] They are used to monitor the process of rehabilitation over a longer temporal context beginning with surgery or compare the impact of different treatments or speech enhancement. Subjective methods involve, for instance, questionnaires about voice-related quality of life[11] and perceptive voice evaluation by speech therapists. Objective methods are based on physical measures, such as frequency analysis or aerodynamic measures. This article focuses on acoustic measures of connected speech because voice and speech parameters are crucial for the ability to communicate.

In clinical routine, different voice and speech criteria are used for perceptive evaluation based on connected words or sentences.[12] Perceptual analysis by speech therapists or naive listeners is still widely used for almost all types of voice disorders.[13–15] However, individual perception may be biased. Averaging over many subjective judgments from different listeners stabilizes the result but is time consuming and not suitable in clinical practice. Additional problems arise from differences in the intra- and interrater correlations as a result of varying test conditions,[16,17] differences in the individual experience, and varying personal conditions of one rater[18,19] or speaker characteristics, such as gender.[20]

In this study, two data sets of speech samples recorded after laryngeal cancer were used. The first set consists of patients who underwent partial laryngectomy (PL) that allowed the preservation of at least one vocal fold. The second set comprises patients with tracheoesophageal (TE) substitute voices after total laryngectomy. Total removal of the larynx has to be performed in 20–40% of all cases of laryngeal or hypopharyngeal cancer.[21] The state-of-the-art voice rehabilitation in these patients is the insertion of shunt valves.[22] During exhalation, the valve redirects the airstream into the upper part of the esophagus.

Accepted for publication April 29, 2011.

From the *Department of Phoniatrics and Pediatric Audiology, University Hospital Erlangen, Erlangen, Germany; and the †Pattern Recognition Lab, Department of Computer Science, University of Erlangen-Nuremberg, Erlangen, Germany.

Address correspondence and reprint requests to Tobias Bocklet, Department of Phoniatrics and Pediatric Audiology, University Hospital Erlangen, Bohlenplatz 21, 91054 Erlangen, Germany. E-mail: tobias.bocklet@informatik.uni-erlangen.de

Journal of Voice, Vol. ■, No. ■, pp. 1-8
0892-1997/$36.00

doi:10.1016/j.jvoice.2011.04.010

TABLE 1.
Statistics of the Two Data Sets Used in This Work

Name	Type	No. of Speakers	Minimum Age (y)	Maximum Age (y)	Mean Age (y)	Standard Deviation of Age
TE	Tracheoesophageal	54	44	84	62.2	10.1
PL	Partial laryngectomy	77	34	83	60.7	9.7

Vibrating tissue of the surrounding pharyngoesophageal segment modulates the flowing airstream and creates the TE voice.[23] This results in lower voice quality, higher strain, and decreased intelligibility.[24]

The acoustic variables of pathologic speech differ from the speech of healthy persons.[24–28] Most work on automatic evaluation of voice aspects focuses on the fundamental frequency (F0) and related features, for example, jitter, shimmer, and length or number of voiced segments, computed on sustained vowels. A reliable automatic F0 extraction is crucial in this case. Voices affected by laryngeal cancer are often aperiodic and contain a high percentage of noise components.[29] F0 extraction algorithms often cannot process highly pathologic voices properly,[30] which makes them unsuitable for fully automatic evaluation systems. Speech properties, such as intelligibility, could not be rated by these approaches because of the missing identification of speech units, such as words or sentences.

We introduce an automatic system based on the evaluation of read speech that can partially solve this problem. The system performs acoustic short-time analysis of a readout text and predicts different voice and speech criteria based on regression. We evaluated the performance of the system on TE and PL speakers with different clinically relevant evaluation criteria, that is, strain, voice quality, and intelligibility.

MATERIALS AND METHODS

Speakers

The statistics of the two different data sets are summarized in Table 1. The speakers of the TE data set were provided with a Provox shunt valve (Atos Medical, Hörby, Sweden).[31] Laryngectomy was performed at least 1 year before recording. Patients with a recurring tumor growth or metastasis were excluded from this study. The speakers of the PL data set had already undergone PL and were recorded on average 2.4 years after surgery.

Stimuli

Each patient was recorded while reading the German version of "The North Wind and the Sun."[32] The text is standardized and phonetically rich, which allows as many different phones as possible to be captured. The recordings were performed with *PEAKS*,[33] a recording and speech analysis tool developed at our working group, and a DNT Call4U headset (DNT, Dietzenbach, Germany) with an attached analog-to-digital converter. Pulse code modulation with a sampling frequency of 16 kHz and 16-bit amplitude resolution was used. The mean duration of the TE speakers' reading of the text was 74 seconds, whereas the PL speakers took approximately 66 seconds to read the text.

Perceptual evaluation

Five speech experts, with at least 5 years experience, participated in the perceptual evaluation experiments. Three different holistic impressions, that is, evaluations regarding the whole utterances, were rated: voice quality, strain, and intelligibility. Voice quality was measured regarding a 10-cm visual analog scale in which the label for a very good voice was at the left end of the scale at position 0.0 cm. Vocal effort and intelligibility were rated on five-point Likert-based[34] scales; the scales of

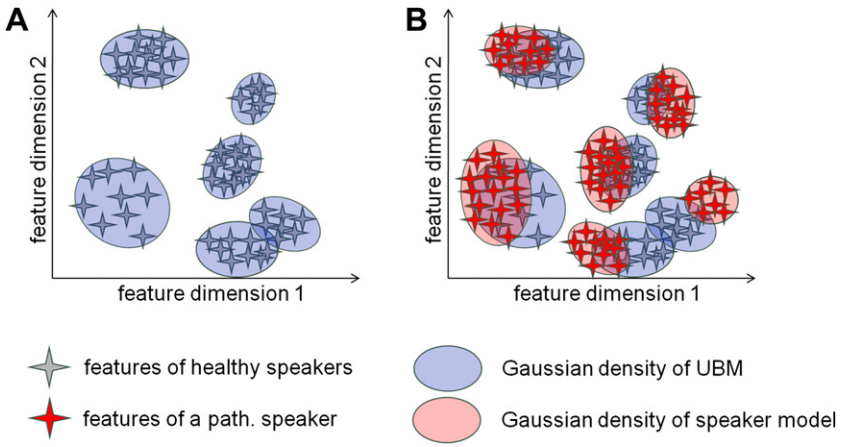

FIGURE 1. **A.** Training of the UBM with healthy speakers using the EM algorithm: The clusters of healthy speakers act as reference (blue). **B.** Adaptation of the UBM to data of a pathologic speaker: The speaker model of a pathologic speaker (red) differs from that of the reference speakers (blue). The distances between corresponding clusters represent the evaluation difference between the pathologic speaker and healthy speakers. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

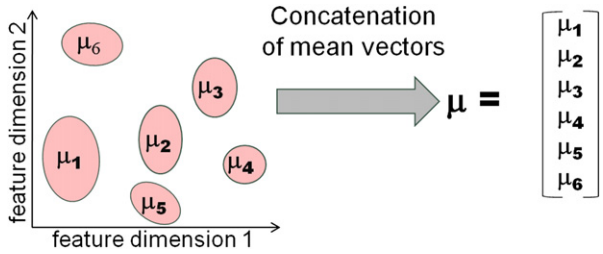

FIGURE 2. Composition of the GMM-based supervector by concatenation of the mean vectors. The supervector concept is an appropriate representation to reduce dimensionality. In this example, the acoustic feature dimension is two and the number of Gaussian densities is six. The final supervector then has a dimension of 12.

these criteria were reciprocal. A "very high" intelligibility rating referring to a better voice was converted to "1" for computation purposes. A "very high" strain was converted to "1" as well. Each rater used the same type of headset (Plantronics .Audio 650 with built-in analog-to-digital converter; Plantronics, Santa Cruz, CA) for listening. In this way, similar evaluation conditions, which were independent of the computer hardware used, could be achieved. The raters of the two data sets differed, which does not allow a direct comparison of the two patient groups in terms of perceptual evaluations. However, this work focuses on an introduction of a completely automatic voice and speech assessment technique and its comparison with perceptual evaluations of human expert listeners. Hence, the system was applied to data of different raters, different data sets, and different voice and speech evaluation criteria to test its ability to measure the degree of voice disorders in patients after laryngeal cancer.

Automatic acoustic modeling

The automatic voice and speech assessment system is based on statistical speaker modeling of the persons' acoustic space. It assumes that the acoustics of pathologic speakers differ from the acoustics of reference speakers without any pathology. The degree of pathology can be measured as distances in the acoustic space between the pathologic speaker model and a reference speaker model. Speakers with a higher degree of pathology have a higher distance to the reference speakers than speakers with a lower degree of pathology.

First, a computational representation of the speech signal must be found. This is a sequence of feature vectors that characterize the speech signal within a short time window. Mel-frequency cepstrum coefficients (MFCCs), feature vectors that are well known in the field of automatic speech and music processing,[35] are used within this work. To compare speakers and their acoustics, modeling from the utterances (with variable length) to speaker-dependent models (with fixed length) must be performed. This is achieved by an unsupervised clustering that projects the MFCCs of speakers to speaker models. These speaker models are statistic representations of the clusters within the acoustic space in terms of Gaussian distributions. These speaker models are then correlated with the different voice and speech evaluation criteria, and prediction models for each of the criteria are trained using support vector regressions (SVR).[36]

Feature extraction: the acoustic space. MFCCs are the standard features in the field of automatic speech processing. These features are based on the frequency perception of humans and perform a frequency analysis of the speech signal. Transformation of the speech signal into the spectral domain is performed by a discrete Fourier transform. The speech signal is decomposed into a series of short stationary segments by a window with a size of 20 milliseconds to account for the typical phoneme duration. The power spectrum is computed afterward. To reduce the number of frequency bands, triangular filters based on the mel scale are used to create 25 spectral coefficients. These coefficients are logarithmic with respect to the loudness perception of the human ear. The cepstral coefficients are computed by an inverse discrete cosine transform of the mel

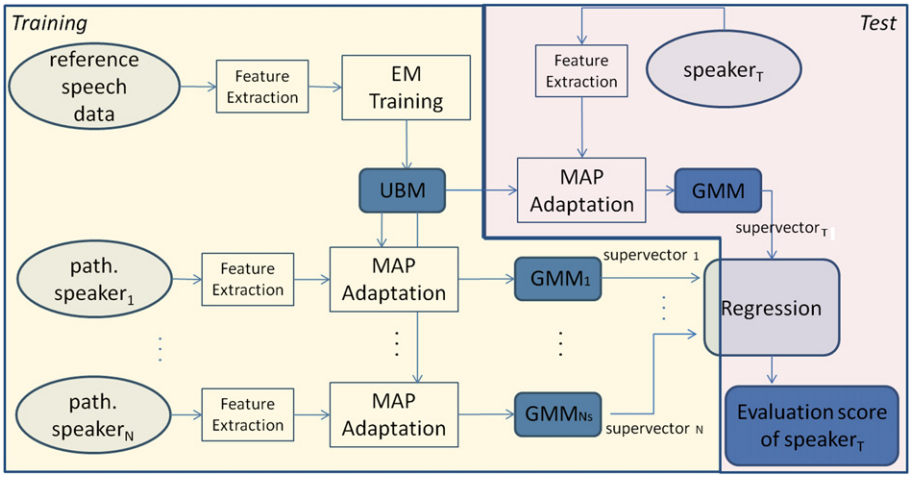

FIGURE 3. Principle of the SVR system. The training sequence is shaded in yellow, and the actual voice testing is shaded in red. In the training sequence, a speaker model (GMM) is created for every training speaker using MAP. The supervectors are extracted, and a regression is trained. In the testing sequence, a GMM for test speaker T is created. The supervector of this speaker is used within the trained regression to evaluate one voice or speech criterion of speaker T. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

TABLE 2.
Interrater Correlation of One Rater to the Mean Value of the Four Remaining Raters for All Rating Criteria on the TE Data Set

	Rater 1		Rater 2		Rater 3		Rater 4		Rater 5		Mean	
Criterion	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
Strain	0.71	0.71	0.80	0.81	0.84	0.85	0.84	0.85	0.73	0.75	0.78	0.79
Voice quality	0.87	0.87	0.87	0.86	0.84	0.83	0.87	0.86	0.80	0.80	0.85	0.84
Intelligibility	0.82	0.83	0.85	0.83	0.81	0.80	0.80	0.80	0.77	0.76	0.81	0.80

The last two columns contain the mean value of the correlations for the single raters.

spectrum coefficients. This step decorrelates the coefficients. The final feature vector is then formed by an extraction of the first 12 coefficients and their first- and second-order derivatives. In this way, a feature vector with dimension $D = 36$ is created.

Speaker modeling. The acoustic features span an acoustic space that is characteristic for a speaker. Similar acoustic features build clusters in the acoustic space. Each cluster represents a different phonetic unit, that is, one particular form of a phoneme. It is defined by a Gaussian distribution in the form of

$$P(c|\mu) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{(1/2)}} e^{-(1/2)(c-\mu)^T \Sigma^{-1}(c-\mu)}$$

where μ denotes the mean vector of the Gaussian density and Σ denotes the covariance. The sum of all acoustic clusters forms a speaker model and can be written as a weighted mixture of the Gaussian densities

$$P(c|\lambda) = \sum_{i=1}^{M} \omega_i p_i(c|\mu_i, \Sigma_i)$$

M denotes the number of clusters, that is, Gaussian densities. ω_i denotes the weights for each density i, $i = 1, \ldots, M$. To create these speaker models, a single speaker-independent Gaussian Mixture Model (GMM) is trained on data of healthy speakers, the so-called Universal Background Model (UBM). This model acts as a reference model of speech uttered by normal nonpathologic voices (Figure 1A). The Gaussian distributions of this model are trained in an unsupervised iterative manner by the Expectation-Maximization algorithm[37] in five iterations. The number of Gaussian densities has to be specified beforehand.

The actual speaker model is derived by adapting the parameters of the UBM using the speech of a pathologic speaker by a kind of Bayesian adaptation, the maximum a posteriori (MAP) adaptation.[38] MAP adapts the density parameters of the UBM to the acoustic feature vectors (MFCCs) of a specific pathologic speaker in a single iteration step. The basic idea in the adaptation approach is to derive the model of the pathologic speaker by updating the well-trained parameters in the UBM (Figure 1B). This results in a speaker-specific GMM with the parameters (ω_i, μ_i, Σ_i, $i = 1, \ldots, M$). To reduce the dimensionality of the speaker model and find a computationally more effective representation, only the mean vectors μ_i are used to represent a speaker. This is achieved by a straightforward concatenation of the M mean vectors μ_i, $i = 1, \ldots, M$, which results in a so-called GMM-based supervector. The extraction sequence of the GMM-based supervectors is depicted in Figure 2.

Regression system for voice and speech parameters. The prediction system is then trained on the GMM-based supervectors created. The system models the interrelation of the GMM-based supervectors and the voice and speech evaluation scores of the human experts. This interrelation can be described by a regression in which one set of variables—the GMM-based supervector—is correlated with another variable, for example, the intelligibility rating of human experts. In the case of a linear regression, this pair of variables is modeled linearly by an affine function. Rather than linear regression, SVR is used for prediction. It depends on a subset of all the data pairs. The cost function for building the model ignores any training data close to the model prediction. For objectivity reasons, the arithmetic means of the perceptual evaluations across all raters are used as "ground truth" for the automatic system. Each evaluation criterion, that is, strain, voice quality, and intelligibility, is described by a different regression. The other parts of the system remain unchanged. The system performs the evaluation in real time. The result is available immediately after recording (Figure 3).

TABLE 3.
Interrater Correlation of One Rater to the Mean Value of the Four Remaining Raters for All Rating Criteria on the PL Data Set

	Rater 1		Rater 2		Rater 3		Rater 4		Rater 5		Mean	
Criterion	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
Strain	0.80	0.81	0.69	0.69	0.85	0.85	0.86	0.86	0.84	0.85	0.81	0.81
Voice quality	0.85	0.86	0.81	0.82	0.88	0.89	0.91	0.90	0.87	0.85	0.86	0.86
Intelligibility	0.82	0.82	0.76	0.76	0.79	0.79	0.86	0.86	0.82	0.84	0.81	0.81

The last two columns contain the mean value of the correlations for the single raters.

TABLE 4.
Human-Machine Correlation of the Automatic System and the Mean Value of All Raters for the Different Criteria on the TE Data Set

Criterion	r	ρ
Strain	0.81	0.80
Voice quality	0.86	0.88
Intelligibility	0.83	0.85

Statistics

A language and environment for statistical computation called R[39] was used for statistical analyses. Pearson's (r) and Spearman's (ρ) correlation coefficients were used to measure the interrelation between different measures. For the comparison of perceptual evaluations, the scores of one rater were compared with the mean value of the four remaining raters. The correlation between the automatic system and the perceptual evaluations was calculated with respect to the average perceptual score of the five raters. Significance tests indicated whether the evaluations of the automatic system and the perceptual evaluations of the expert listeners were notably different.

RESULTS

Perceptual evaluations

Tables 2 and 3 show the interrater correlation for all rating criteria on the TE and PL data set, respectively. Each rater was compared with the mean value of the remaining raters. None of the differences between the rater correlations were significant ($P > 0.1$).

Automatic acoustic modeling

The correlations between the automatic system and the average perceptual analysis on the TE and PL groups are summarized in Tables 4 and 5, respectively. Figures 4–6 show the correlation for each criterion in detail for TE speakers; Figures 7–9 present it for the PL speakers.

Experiments in this study did not show significant differences at $P < 0.05$ when different numbers of Gaussian densities were used for speaker modeling. For this reason, all results are reported for speaker models with 256 Gaussian densities.

DISCUSSION

An automatic method for acoustic speaker modeling was applied to three voice- and speech-related evaluation criteria:

TABLE 5.
Human-Machine Correlation of the Automatic System and the Mean Value of All Raters for the Different Criteria on the PL Data Set

Criterion	r	ρ
Strain	0.80	0.79
Voice quality	0.84	0.84
Intelligibility	0.82	0.82

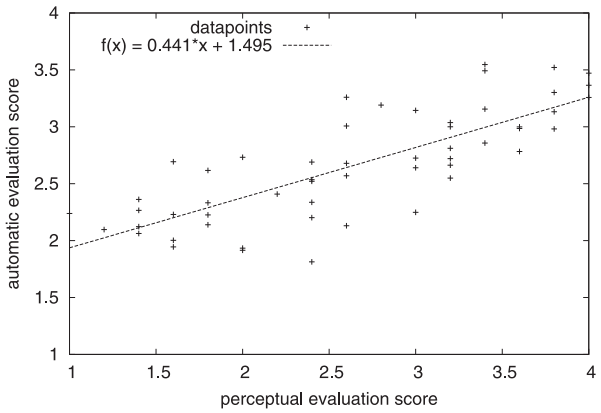

FIGURE 4. Strain evaluations: perceptual versus automatic scores (TE; $r = 0.81$; $\rho = 0.80$). The x-axis refers to the perceptual evaluations (mean value of five raters), and the y-axis denotes the scores of the automatic system. The dotted line denotes the linear regression line between perceptual and automatic evaluations.

strain, voice quality, and intelligibility. The system has been evaluated on two different data sets of speakers after laryngeal cancer. The ground truth for the automatic evaluations was the perceptual evaluations by expert listeners. Perceptual evaluations of different raters are never completely consistent; raters always deviate to a certain degree. This holds for intra- and interrater consistency. Nevertheless, perceptual evaluations are still a widely used standard technique for voice and speech evaluations. To compensate for the differences in evaluations, the arithmetic mean value of the five expert raters was used. The problem of intrarater variability does not apply for automatic evaluation systems because one specific recording always achieves the same result when evaluated several times.

We intentionally did not use Cronbach's α[40] or Cohen's κ[41] for interrater agreement measurement because α and κ are defined for integer values only. The arithmetic mean of the raters' scores and the real-valued automatic measures, however, are continuous values. To allow a fair comparison between the

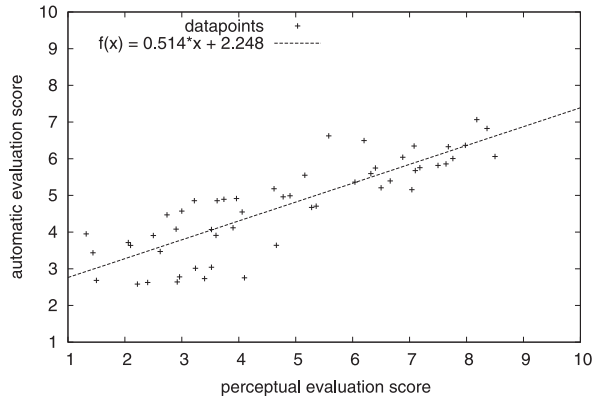

FIGURE 5. Voice quality evaluations: perceptual versus automatic scores (TE; $r = 0.86$; $\rho = 0.88$). The x-axis refers to the perceptual evaluations (mean value of five raters), and the y-axis denotes the scores of the automatic system. The dotted line denotes the linear regression line between perceptual and automatic evaluations.

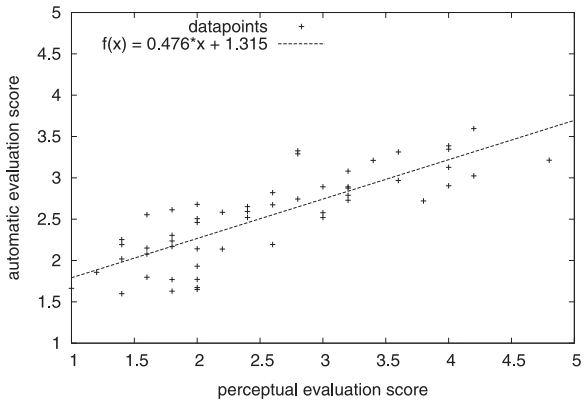

FIGURE 6. Intelligibility evaluations: perceptual versus automatic scores (TE; $r = 0.83$; $\rho = 0.85$). The x-axis refers to the perceptual evaluations (mean value of five raters), and the y-axis denotes the scores of the automatic system. The dotted line denotes the linear regression line between perceptual and automatic evaluations.

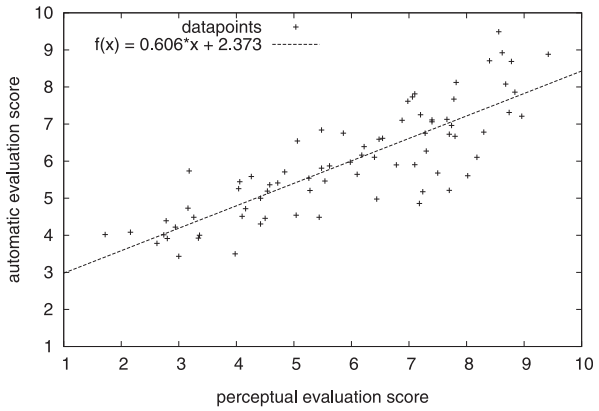

FIGURE 8. Voice quality evaluations: perceptual versus automatic scores (PL; $r = 0.84$; $\rho = 0.84$). The x-axis refers to the perceptual evaluations (mean value of five raters), and the y-axis denotes the scores of the automatic system. The dotted line denotes the linear regression line between perceptual and automatic evaluations.

agreement of the human expert raters and the automatic system, Pearson's and Spearman's correlation coefficients were calculated between one rater and the mean value of the four remaining raters. The interrater results in Tables 2 and 3 show high average correlations (r, $\rho \geq 0.78$) for the different evaluation criteria on the two data sets. However, there is some divergence for some specific raters. Note that the perceptual evaluations were performed by speech experts of the same department. Studies focusing on perceptual evaluations show even higher interrater variabilities.[13,42,43] In a clinical environment, this is a problem when only one expert's opinion can be obtained. Measuring the progress of a therapy over a longer temporal context requires evaluation of the same expert listener, which is not always possible.

For both data sets, the correlation between the automatic system and the perceptual evaluations is r, $\rho \geq 0.79$ (Tables 4 and 5). These correlations are in the same range as the expert ratings; the differences are not significant. The usability of the system has been shown by high correlations for different data sets, different raters, and different evaluation criteria. This holds for vocal parameters, such as voice quality and vocal effort, and for speech parameters, such as intelligibility.

The proposed system evaluates a speaker by a statistical model based on short-time acoustic analysis. If the acoustic factors of influence, for example, background noise, type of microphone, or spoken text, are kept constant for each recording, the degrees of freedom of the speaker models are reduced to the identity of the speaker and his/her voice and speech characteristics. In the training of the system's regression component, the system learns which parameters to rely on for evaluation. The speaker-dependent factors of the speaker model are excluded implicitly in the training step of the regression component.

The system evaluates a spoken text rather than sustained vowels. Automatic evaluations on sustained vowels often focus on an analysis of the F0 and its variations in time and amplitude (jitter and shimmer). The problems regarding a robust F0

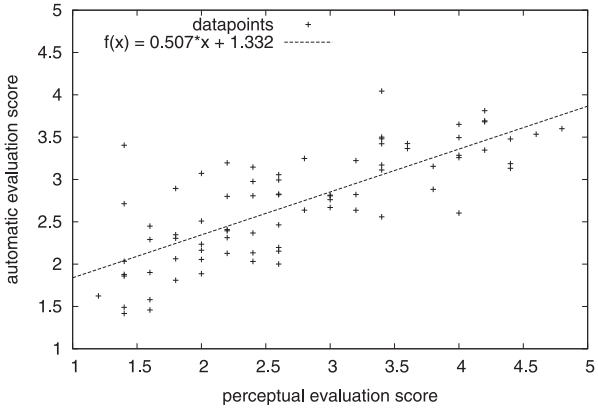

FIGURE 7. Strain evaluations: perceptual versus automatic scores (PL; $r = 0.80$; $\rho = 0.79$). The x-axis refers to the perceptual evaluations (mean value of five raters), and the y-axis denotes the scores of the automatic system. The dotted line denotes the linear regression line between perceptual and automatic evaluations.

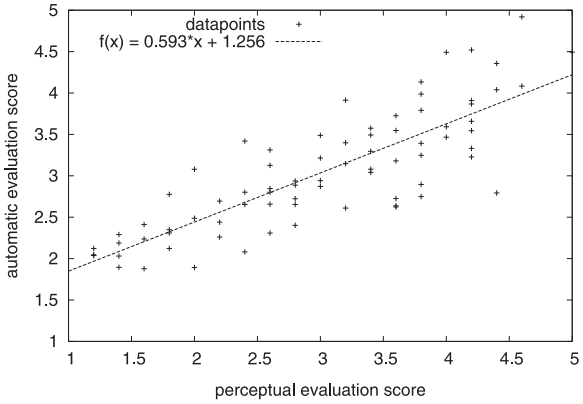

FIGURE 9. Intelligibility evaluations: perceptual versus automatic scores (PL; $r = 0.82$; $\rho = 0.82$). The x-axis refers to the perceptual evaluations (mean value of five raters), and the y-axis denotes the scores of the automatic system. The dotted line denotes the linear regression line between perceptual and automatic evaluations.

extraction for highly pathologic voices[30] do not apply with the proposed system. The fact that the speaker models contain acoustic information without temporal order basically implies that the temporal information is not taken into account for the evaluations. The system rather focuses on different phonetic subclasses represented by the Gaussian densities. This makes the system robust against reading errors and allows a more detailed automatic analysis of different phonemes. Future work will focus on this aspect.

CONCLUSION

This article introduced a novel automatic system for evaluations of acoustic parameters of connected speech, which are one of the aspects for voice and speech evaluation. The results of the analysis are available immediately after the speech recording has been obtained. The method is not biased by varying intrarater correlations and always produces the same evaluation score for identical recordings. Further applications, such as interactive training tools for home use for affected persons, are also possible with an adequate graphical user interface. The approach is not intended to replace perceptual evaluations. Instead, it can be used in clinical practice as objective support for therapy outcome assessment and therapy control. It can act as a second objective opinion for acoustic analysis when multiple listeners are not available or too expensive. This saves time and money, and the speech therapists can spend more time on interacting with the patients.

Acknowledgments

This study was partially funded by the German Cancer Aid (Deutsche Krebshilfe) under grant 107873 and by the German Research Foundation (DFG) under grant EY 15/18-2. We thank the speech therapists of our department for the perceptual evaluations.

REFERENCES

1. Fritzell B. Voice disorders and occupations. *Logoped Phoniatr Vocol*. 1996; 21:7–12.
2. Titze IR, Lemke J, Montequin D. Population in the U.S. workforce who rely on voice as a primary tool of trade: a preliminary report. *J Voice*. 1997;11: 254–259.
3. Ruben R. Redefining the survival of the fittest: communication disorders in the 21st century. *Laryngoscope*. 2000;110:241–245.
4. Hummel C, Scharf M, Schuetzenberger A, Graessel E, Rosanowski F. Objective voice parameters and self-perceived handicap in dysphonia. *Folia Phoniatr Logop*. 2010;62:303–307.
5. Krischke S, Weigelt S, Hoppe U, Köllner V, Klotz M, Eysholdt U, Rosanowski F. Quality of life in dysphonic patients. *J Voice*. 2005;19: 132–137.
6. Branski RC, Cukier-Blaj S, Pusic A, et al. Measuring quality of life in dysphonic patients: a systematic review of content development in patient-reported outcomes measures. *J Voice*. 2010;24:193–198.
7. Kazi R, De Cordova J, Singh A, et al. Voice-related quality of life in laryngectomees: assessment using the VHI and V-RQOL symptom scales. *J Voice*. 2007;21:728–734.
8. Harrison AE. *Speech Disorders: Causes, Treatment and Social Effects*. Hauppauge NY: Nova Science Publishers; 2010.
9. Devins GM, Stam HJ, Koopmans JP. Psychosocial impact of laryngectomy mediated by perceived stigma and illness intrusiveness. *Can J Psychiatry*. 1994;39:608–616.
10. Dejonckere PH, Bradley P, Clemente P, et al, 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.
11. Hogikyan ND, Sethuraman G. Validation of an instrument to measure voice-related quality of life (V-RQOL). *J Voice*. 1999;13:557–569.
12. Schiavetti N. Scaling procedures for the measurement of speech intelligibility. In: Kent RD, ed. *Intelligibility in Speech Disorders: Theory, Measurement and Management*. Philadelphia, PA: John Benjamins; 1992: 11–34.
13. van As CJ, Koopmans-van Beinum FJ, Pols LCW, Hilgers FJM. 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.
14. Ainsworth WA, Singh W. Perceptual comparison of neoglottal, oesophageal and normal speech. *Folia Phoniatr (Basel)*. 1992;44:297–307.
15. Wolfe VI, Martin DP, Palmer CI. Perception of dysphonic voice quality by naive listeners. *J Speech Lang Hear Res*. 2000;43:697–705.
16. McColl DA. Intelligibility of tracheoesophageal speech in noise. *J Voice*. 2006;20:605–615.
17. McColl D, Fucci D, Petrosino L, Martin DE, Mc Caffrey P. Listener ratings of the intelligibility of tracheoesophageal speech in noise. *J Commun Disord*. 1998;31:279–289.
18. Bunton K, Kent RD, Duffy JR, Rosenbek JC, Kent JF. Listener agreement for auditory-perceptual ratings of dysarthria. *J Speech Lang Hear Res*. 2007;50:1481–1495.
19. Sheard C, Adams RD. Reliability and agreement of ratings of ataxic dysarthric speech samples with varying intelligibility. *J Speech Hear Res*. 1991;34:285–293.
20. Eadie TL, Doyle PC, Hansen K, Beaudin PG. Influence of speaker gender on listener judgments of tracheoesophageal speech. *J Voice*. 2008;22:43–57.
21. van der Torn M, Mahieu HF, Festen JM. Aero-acoustics of silicone rubber lip reeds for alternative voice production in laryngectomees. *J Acoust Soc Am*. 2001;110:2548–2559.
22. 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.
23. Schutte HK, Nieboer GJ. Aerodynamics of esophageal voice production with and without a Groningen voice prosthesis. *Folia Phoniatr Logop*. 2002;54:8–18.
24. Moerman M, Pieters G, Martens JP. Objective evaluation of the quality of substitution voices. *Eur Arch Otorhinolaryngol*. 2004;261:541–547.
25. van As CJ, Hilgers FJM, Verdonck-de Leeuw IM, Koopmans-van Beinum FJ. Acoustical analysis and perceptual evaluation of tracheoesophageal prosthetic voice. *J Voice*. 1998;12:239–248.
26. MacCallum JK, Cai L, Zhou L, Zhang Y, Jiang JJ. Acoustic analysis of aperiodic voice: perturbation and nonlinear dynamic properties in esophageal phonation. *J Voice*. 2009;23:283–290.
27. Torrejano G, Guimaraes I. Voice quality after supracricoid laryngectomy and total laryngectomy with insertion of voice prosthesis. *J Voice*. 2009; 23:240–246.
28. Zhang Y, Jiang JJ. Acoustic analyses of sustained and running voices from patients with laryngeal pathologies. *J Voice*. 2008;22:1–9.
29. 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.
30. Bocklet T, Toy H, Nöth E, Schuster M, Eysholdt U, Rosanowski F, Gottwald F, Haderlein T. Automatic evaluation of tracheoesophageal substitute voice: sustained vowel versus standard text. *Folia Phoniatr Logop*. 2009;61:112–116.
31. Hilgers FJM, Schouwenburg PF. A new low-resistance, self-retaining prosthesis (Provox®) for voice rehabilitation after total laryngectomy. *Laryngoscope*. 1990;100:1202–1207.
32. *Handbook of the International Phonetic Association*. Cambridge, UK: Cambridge University Press; 1999.

33. Maier A, Haderlein T, Eysholdt U, Rosanowski F, Batliner A, Schuster M, Nöth E. PEAKS—a system for the automatic evaluation of voice and speech disorders. *Speech Commun*. 2009;51:425–437.
34. Likert R. *A Technique for the Measurement of Attitudes. Archives of Psychology*, Vol 140. New York, NY: Columbia University; 1932: 1–55.
35. 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.
36. Smola AJ, Schölkopf B. A tutorial on support vector regression. *Stat Comput*. 2004;14:199–222.
37. Dempster AP, Laird NM, Rubin DB. Maximum-likelihood from incomplete data via the EM algorithm. *J R Stat Soc Series B (Methodol)*. 1977;39:1–38.
38. Gauvain JL, Lee CH. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov Chains. *IEEE Trans Speech Audio Process*. 1994;2:291–298.
39. R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2008. Available at: http://www.R-project.org. Accessed May 27, 2011.
40. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika*. 1951;16:297–334.
41. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;XX:37–46.
42. Dejonckere PH, Remacle M, Fresnel-Elbaz E, Woisard V, Crevier-Buchman L, Millet B. Differentiated perceptual evaluation of pathological voice quality: reliability and correlations with acoustic measurement. *Rev Laryngol Otol Rhinol (Bord)*. 1996;117:219–224.
43. Kreiman J, Gerratt BR, Kempster GB, Erman A, Berke GS. Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. *J Speech Hear Res*. 1993;36:21–40.