Text-based vs. Vowel-based Automatic Evaluation of Tracheoesophageal Substitute Voice

Tino Haderlein[1,2], Tobias Bocklet[1,2], Elmar Nöth[2], Frank Rosanowski[1]
[1]Department of Phoniatrics and Pedaudiology, University of Erlangen-Nuremberg
Bohlenplatz 21, 91054 Erlangen, Germany
Phone: +49 9131 852-7825, Fax: +49 9131 303811, E-Mail: Tino.Haderlein@informatik.uni-erlangen.de
[2]Chair of Pattern Recognition (Computer Science 5), University of Erlangen-Nuremberg
Martensstraße 3, 91058 Erlangen, Germany

Keywords: Substitute voice, Automatic speech recognition, Hoarseness Diagram, Prosodic features

Abstract – The Hoarseness Diagram, a program for voice quality analysis using recordings of sustained vowels, was compared to an automatic speech recognition system with a module for prosodic analysis. The latter computed prosodic features on a text recording. We examined whether the voice analysis of sustained vowel and text analysis correlate on a group of 24 male laryngectomees (average age: 60.6±8.9 years) using tracheoesophageal substitute speech. Each person read the German version of the text "The North Wind and the Sun" which consists of 108 words. Additionally, 5 sustained vowels were recorded from each patient.

The correlation between the measures obtained by the Hoarseness Diagram and the prosodic features from the prosody module was determined. Parameters like jitter, shimmer, F0 and irregularity computed by the Hoarseness Diagram on vowel recordings show correlations of about –0.8 to prosodic features obtained from the text recordings. Hence, voice properties can reliably be evaluated both on a vowel and a text recording. The text analysis, however, offers also possibilities for automatic speech evaluation since it represents a real communication situation better.

1. INTRODUCTION

Laryngectomy, i.e. removal of the larynx, because of laryngeal or hypopharyngeal cancer affects many aspects of life with loss of ability for vocal communication being an outstanding stigma for the affected persons. The tracheoesophageal (TE) substitute voice is state-of-the-art of voice rehabilitation after laryngectomy [1]: A silicone one-way valve is placed into a shunt between the trachea and the esophagus which prevents aspiration and deviates the air stream during expiration into the upper esophagus. The upper esophagus, the pharyngo-esophageal segment (PE segment; see Fig. 1), serves as substitute sound generator. Tissue vibrations of the PE segment modulate the streaming air and generate the primary substitute voice signal which is then further modulated in the same way as normal speech. The quality of substitute voice is "low" when compared to normal voice [2,3] with a loss of prosodic features being one particular characteristic.

When a patient's voice has to be evaluated for clinical purposes, evaluation is mainly performed by human raters. This may be biased and is time-consuming. Automatically computed objective measures are helpful since they provide a solution for these two problems: Costs of human resources are reduced and the problem of inter- and intra-rater variability is eliminated because an automated evaluation algorithm always yields the same result for one specific audio recording.

The Hoarseness Diagram (HD, [4]) is an established method for rating the quality of pathologic voice in German-speaking countries. It is mainly used for evaluation of laryngeal hoarseness and processes recordings of sustained vowels. However, a sustained vowel does not reflect the patients' everyday communication. This is why the analysis of free speech or a read out text should be preferred for evaluation. In earlier publications, we showed that evaluation criteria like intelligibility, speaking effort, or the match of breath and sense units can be obtained automatically from a read standard text [5,6]. Technically this is achieved by a speech recognition system and a "prosody module". Prosodic information is attached to speech segments above phoneme level, i.e. syllables, words, phrases, and whole utterances. Perceptive properties like pitch, loudness, articulation rate, voice quality, duration, pause or rhythm are assigned to these segments. The prosody module computes features based upon frequency, duration and speech energy measures.

In this paper, the correlation between the vowel-based analysis of the HD and the text-based analysis of the prosody module will be presented.

2. SOFTWARE

2.1 The Hoarseness Diagram

The Hoarseness Diagram (distributed by Rehder/Partner, Hamburg, Germany) is a product for analyzing voice quality with respect to hoarseness. It is clinically validated and has been in use in German-speaking countries for several years [4]. It provides a two-dimensional graphical representation of voice quality. The x-axis represents the degree of irregularity, and the y-axis shows the noisy fraction of the voice. For the calculation of these characteristics, sustained vowels are recorded with a specific microphone provided with the program. The computation of irregularity ("irreg") is based on three acoustic features: jitter (variation of period length), shimmer (variation of energy) and the short-time cross-

correlation of adjacent cyclic periods ("p-corr"). These features are specific for the roughness of voice; the also provided "noise" component measures the amount of pulselike vs. noisy voice excitation and is based only on the Glottal to Noise Excitation Ratio (GNE, [7]). It is independent from jitter and shimmer and expresses how voicing is excited by glottal activity or turbulent noise. Hence, it is a measure for breathiness. The more the voice is perturbed, the further away it is plotted from the point of origin of the 2-D diagram.

2.2 The Recognition System and the Prosody Module

The speech recognition system used for the experiments was developed at the Chair of Pattern Recognition in Erlangen. It can handle spontaneous speech with mid-sized vocabularies up to 10,000 words. The latest version is described in detail in [8]. The system is based on semi-continuous Hidden Markov Models (HMM). It can model phones in a context as large as statistically useful and thus forms the so-called polyphones, a generalization of the well-known bi- or triphones [9]. The HMMs for each polyphone have three to four states; the codebook has 500 classes with full covariance matrices. The short-time analysis applies a Hamming window with a length of 16 ms, the frame rate is 10 ms. The filterbank for the Mel-spectrum consists of 25 triangle filters. For each frame, a 24-dimensional feature vector is computed. It contains short-time energy, 11 Mel-frequency cepstrum coefficients (MFCC, [10]), and the first-order derivatives of these 12 static features. The derivatives are approximated by the slope of a linear regression line over 5 consecutive frames (56 ms). Only a unigram language model was used so that the results are mainly dependent on the acoustic models.

The system for the experiments in this paper was trained with German dialogues from the Verbmobil project [11]. The data were recorded with a close-talking microphone (16 kHz, 16 bit). The 578 training speakers (304 male, 274 female) were from all over Germany. About 80% of them were between 20 and 29 years old, less than 10% were over 40. 11,714 utterances (257,810 words) of the Verbmobil-German data (27.7 hours of speech) were used for training and 48 (1042 words) for the validation set [8].

The vocabulary of the system was changed to the 71 words of the text "Der Nordwind und die Sonne", a phonetically rich text with 108 words (71 disjunctive) and 172 syllables. It is used in medical speech evaluation in German-speaking countries and was read by our test speakers. Its English version is known as "The North Wind and the Sun" [12].

The prosody module for the analysis of the standard text requires a word hypotheses graph (WHG) from the recognition system as input which contains the information where each word begins and ends in the respective recording. This time-alignment was done by the speech recognition module on a word-wise transliteration of the spoken text.

The prosody module derives 95 "local" features for each processed word and 15 "global" features per recording, i.e. on the entire text. In this paper, the focus is on the global features since they proved to be more suitable for the task than the local features in pilot experiments (unpublished data). The global features are based on jitter, shimmer, and the number of voiced/unvoiced sections in the speech signal. Among them are the mean value and standard deviation of jitter and shimmer, the number, length and maximum length of voiced and unvoiced speech 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 recording, and the same for unvoiced sections. The last global feature is the standard deviation of the F0. The decision whether a section is voiced or not is based upon the signal intensity which is higher during voicing and on the zero crossing rate of the amplitude which is usually low for a voiced signal. More details and further references concerning the features are given in [13].

3. TEST SPEAKERS AND SAMPLES

Audio files were recorded from 24 male laryngectomees with tracheoesophageal substitute voice. Their average age was 60.6 years (standard deviation: 8.9 years). All of them had been provided with a Provox® shunt valve.

The samples of the text "Der Nordwind und die Sonne" were recorded with a close-talking microphone (dnt Call 4U Comfort headset; DNT, Dietzenbach, Germany) and digitized with 16 bit at 16 kHz sampling frequency. All recordings were made in a small, quiet room with clinical routine acoustic properties. The duration of all 24 audio files together was 29.5 minutes (average: 74.5 seconds per speaker, standard dev.: 27.4 seconds), the test persons spoke 2637 words (average: 109.9 words per speaker, standard dev.: 2.4 words). In addition to the words of the text reference, they produced 38 different additional words and word fragments (45 in total) due to reading errors. In the same room, 5 sustained vowels (/a/, /e/, /i/, /o/, /u/) were recorded from each patient with the HD microphone.

4. RESULTS

Table 1 shows Pearson's correlation between the measures calculated by the Hoarseness Diagram on a sustained vowel and those obtained by the prosody module on the standard text. The table contains all features where the correlation was $|r| \geq 0.7$ for at least one vowel. Correlations with $|r| \geq 0.7$ between the mean values of all vowels on the HD and the prosody module could be observed for the HD features irregularity, jitter, F0, shimmer and mean waveform correlation. The HD features GNE and noise correlated with the prosodic features only with $|r| \leq 0.5$ and were therefore not further examined.

5. DISCUSSION

Usually, pathologic voice quality is evaluated automatically on recordings of sustained vowels only. For hoarseness, however, it has been confirmed that acoustic parameters from connected speech are more reliable than those of sustained vowels [14]. For tracheoesophageal speech, it was shown that automatic analysis of prosodic features on a read out text shows strong correlation to human evaluation criteria, like "intelligibility", "speaking effort" or "match of breath and sense units" [6]. In this

paper, we examined whether prosodic features obtained on a standard text correlate to the measures that are computed by the Hoarseness Diagram which is an established approach for vowel analysis.

For the calculation of jitter and "irregularity" in general, the fundamental frequency (F0) plays an important role. In recordings of substitute voice, it is usually difficult to extract periodic sections at all. The F0 detection algorithm of the prosody module could not compute the frequency for all of the recordings (vowel 'a': 6, 'e' and 'i': 1, 'u': 2 cases) while the HD always returned a value. However, the detected values showed clearly that they were sometimes outside a valid range. The average male TE speaker has an average F0 of below 100 Hz [3]. Automatically computed frequencies of 200 or even 400 Hz are very likely caused by octave errors. The Hoarseness Diagram produces many of these errors, especially in those cases where the prosody module decided for the absence of a voiced signal and returned nothing.

The measures jitter, shimmer and irregularity of the HD correlate with some of the global prosodic duration features very well ($|r| \geq 0.7$). Especially measures like the maximum length of the voiced or unvoiced section in a word, the ratio of the length of voiced and unvoiced sections, or the total length of voiced sections in the entire text correspond with the HD measures of a single vowel. It is obvious that the correlation for all these features is negative because the more irregular a voice is, the shorter are the voiced sections in speech. In a highly irregular voice, jitter and shimmer are much higher and – in the case of the HD's F0 detection algorithm – also the values for F0. When comparing these F0 values of the HD to the F0-based prosodic features obtained on text recordings, no correlations of $r \geq 0.7$ were found. The correlation between the F0 values of HD and prosody module was $r \leq 0.36$ even when the cases where the prosody module returned 0 for the F0, i.e. "unvoiced", were excluded.

For almost all features, the best correlation is achieved when the average value of the respective HD measure on all 5 vowels is compared to the specific prosodic feature. In order to achieve a good correlation for all the HD measures, it is obviously necessary to record all these vowels. Although most of the HD measures are highly correlated to prosodic features, no vowel reached $|r| \geq 0.5$ for the HD measures GNE and noise which represent the two axes in the graphical output of the program. The combination of prosodic features to match GNE and noise better has not been examined yet.

It was shown that the human evaluation criterion "match of breath and sense units" correlates with several duration and pause features of the prosody module very well ($r > 0.8$; [6]). There is no sufficient correlation between HD features and pause features of the prosody module. This means that the individual speaking properties, like the rhythm of breathing and the articulation rate, cannot be determined by a method that only analyzes sustained vowels. The fact that the pause features are very important for automatic speech evaluation leads to the conclusion that voice pathology should be evaluated by means of a full read out text and not only on isolated vowel recordings. Nevertheless, the text analysis cannot replace the vowel analysis completely because the prosody module computes averaged features for a text that contains many different vowels and consonants.

6. CONCLUSION

Individual speaking properties cannot be determined by a method that only analyzes sustained vowels. A text-based evaluation, however, computes features that are averaged across many different phones which might cause the loss of interesting information. We therefore suggest that automatic evaluation of speech pathology should be performed on both a sustained vowel and a text in order to cover the properties of a patient's voice and speech as good as possible.

ACKNOWLEDGMENTS

This work was partially funded by the German Cancer Aid (Deutsche Krebshilfe, grant 106266). The responsibility for the contents of this study lies with the authors.

REFERENCES

[1] D.H. Brown, F.J.M. Hilgers, J.C. Irish, A.J.M. Balm, "Postlaryngectomy Voice Rehabilitation: State of the Art at the Millennium", World J. Surg., vol. 27, pp. 824–831, 2003.

[2] M.H. Bellandese, J.W. Lerman, H.R. Gilbert, "An Acoustic Analysis of Excellent Female Esophageal, Tracheoesophageal, and Laryngeal Speakers", J. Speech Lang. Hear. Res., vol. 44, pp. 1315–1320, 2001.

[3] J. Robbins, H.B. Fisher, E.D. Blom, M.I. Singer, "A Comparative Acoustic Study of Normal, Esophageal, and Tracheoesophageal Speech Production", J. Speech Hear. Disord., vol. 49, pp. 202–210, 1984.

[4] M. Fröhlich, D. Michaelis, H.W. Strube, E. Kruse, "Acoustic voice analysis by means of the hoarseness diagram", J. Speech Lang. Hear. Res., vol. 43, pp. 706–720, 2000.

[5] M. Schuster, T. Haderlein, E. Nöth, J. Lohscheller, U. Eysholdt, F. Rosanowski, "Intelligibility of Laryngectomees' Substitute Speech: Automatic Speech Recognition and Subjective Rating", Eur. Arch. Otorhinolaryngol., vol. 263, pp. 188–193, 2006.

[6] T. Haderlein, E. Nöth, H. Toy, A. Batliner, M. Schuster, U. Eysholdt, J. Hornegger, F. Rosanowski, "Automatic Evaluation of Prosodic Features of Tracheoesophageal Substitute Voice", Eur. Arch. Otorhinolaryngol., vol. 264, pp. 1315–1321, 2007.

[7] D. Michaelis, T. Gramss, H.-W. Strube, "Glottal to noise excitation ratio – a new measure for describing pathological voices", Acustica/acta acustica, vol. 83, pp. 700–706, 1997.

[8] G. Stemmer, *Modeling Variability in Speech Recognition*, vol. 19 of Studien zur Mustererkennung. Logos Verlag, Berlin, 2005.

[9] E.G. Schukat-Talamazzini, H. Niemann, W. Eckert, T. Kuhn, S. Rieck, "Automatic Speech Recognition without Phonemes", in *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*, European Speech Communication Association (ESCA), Berlin, pp. 129–132, 1993.

[10] S.B. Davis, P. Mermelstein, "Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Trans. on Acoustics, Speech and Signal Processing, vol. 28, pp. 357–366, 1980.

[11] W. Wahlster, ed., *Verbmobil: Foundations of Speech-to-Speech Translation*, Springer, Berlin, 2000.
[12] International Phonetic Association, *Handbook of the International Phonetic Association*, Cambridge University Press, Cambridge, 1999.
[13] A. Batliner, K. Fischer, R. Huber, J. Spilker, E. Nöth, "How to Find Trouble in Communication", Speech Communication, vol. 40, pp. 117–143, 2003.
[14] B. Halberstam, "Acoustic and Perceptual Parameters Relating to Connected Speech Are More Reliable Measures of Hoarseness than Parameters Relating to Sustained Vowels", ORL, vol. 66, pp. 70–73, 2004.
[15] J. Lohscheller, *Dynamics of the Laryngectomee Substitute Voice Production*, vol. 14 of Berichte aus Phoniatrie und Pädaudiologie, Shaker Verlag, Aachen, 2003.

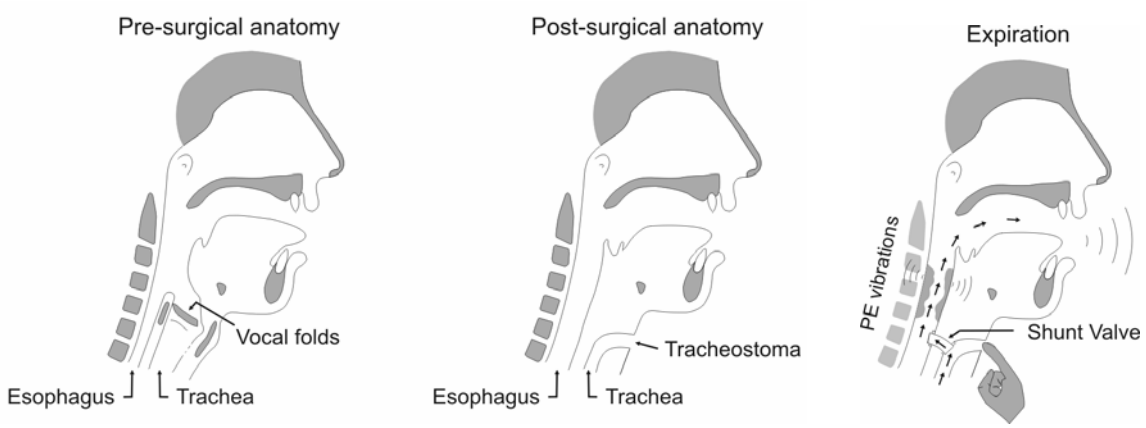

Figure 1: Anatomy of a person with intact larynx (left), anatomy after total laryngectomy (middle), and the substitute voice (right) caused by vibration of the pharyngoesophageal segment (pictures from [15])

HD	prosodic feature	a	e	i	o	u	avg.
irreg	maximum length of voiced section	–0.63	–0.67	–0.59	–0.84	–0.69	–0.82
irreg	total length of voiced sections	–0.58	–0.63	–0.60	–0.79	–0.68	–0.79
irreg	ratio length voiced/unvoiced sections	–0.59	–0.64	–0.59	–0.74	–0.70	–0.79
irreg	ratio length voiced sections/length of recording	–0.59	–0.56	–0.64	–0.67	–0.63	–0.74
jitter	maximum length of voiced section	–0.68	–0.64	–0.57	–0.75	–0.67	–0.79
jitter	ratio length voiced/unvoiced sections	–0.65	–0.61	–0.58	–0.65	–0.68	–0.76
jitter	total length of voiced sections	–0.62	–0.61	–0.56	–0.69	–0.66	–0.76
jitter	ratio length voiced sections/length of recording	–0.60	–0.49	–0.70	–0.65	–0.59	–0.73
F0	maximum length of voiced section	–0.78	–0.63	–0.22	–0.63	–0.48	–0.79
F0	ratio length voiced/unvoiced sections	–0.71	–0.59	–0.24	–0.63	–0.54	–0.77
F0	total length of voiced sections	–0.72	–0.58	–0.24	–0.63	–0.50	–0.76
shimmer	maximum length of voiced section	–0.78	–0.70	–0.54	–0.77	–0.66	–0.81
shimmer	ratio length voiced/unvoiced sections	–0.71	–0.66	–0.58	–0.71	–0.68	–0.78
shimmer	total length of voiced sections	–0.71	–0.64	–0.54	–0.73	–0.64	–0.76
shimmer	ratio length voiced sections/length of recording	–0.68	–0.63	–0.59	–0.68	–0.64	–0.75
p-corr	ratio length voiced/unvoiced sections	+0.33	+0.48	+0.64	+0.68	+0.69	+0.70
p-corr	maximum length of voiced section	+0.35	+0.49	+0.61	+0.74	+0.65	+0.71

Table 1: Pearson's correlation between the measures obtained by the Hoarseness Diagram (HD) on recordings of different vowels ("avg." is the average across all vowels) and prosodic features obtained by the prosody module on the entire text. Given are all measure–feature pairs where for at least one vowel the correlation was $|r| \geq 0.7$.