Automatic Evaluation of Tracheoesophageal Substitute Voice: Sustained Vowel versus Standard Text

Tobias Bocklet, Hikmet Toy, Elmar Nöth, Maria Schuster, Ulrich Eysholdt, Frank Rosanowski, Frank Gottwald, Tino Haderlein

Department of Phoniatrics and Pedaudiology, Chair of Pattern Recognition (Computer Science 5), and Department of Otorhinolaryngology – Head and Neck Surgery, University of Erlangen-Nuremberg, Erlangen, Germany

Key Words
Laryngectomy · Substitute voice · Automatic speech recognition · Hoarseness Diagram

Abstract
Objective: The Hoarseness Diagram, a program for voice quality analysis used in German-speaking countries, was compared with an automatic speech recognition system with a module for prosodic analysis. The latter computed prosodic features on the basis of a text recording. We examined whether voice analysis of sustained vowels and text analysis correlate in tracheoesophageal speakers. Patients and Methods: Test speakers were 24 male laryngectomees with tracheoesophageal substitute speech, age 60.6 ± 8.9 years. Each person read the German version of the text ‘The North Wind and the Sun’. Additionally, five sustained vowels were recorded from each patient. The fundamental frequency (F0) detected by both programs was compared for all vowels. The correlation between the measures obtained by the Hoarseness Diagram and the features from the prosody module was computed. Results: Both programs have problems in determining the F0 of highly pathologic voices. Parameters like jitter, shimmer, F0, and irregularity as computed by the Hoarseness Diagram from vowels show correlations of about −0.8 with prosodic features obtained from the text recordings. Conclusion: Voice properties can reliably be evaluated both on the basis of vowel and text recordings. Text analysis, however, also offers possibilities for the automatic evaluation of running speech since it realistically represents everyday speech.
Olthoff et al. [10] reported that it is also a suitable approach for speakers after partial and even total laryngectomy.

Like other automatic evaluation methods [11, 12], the HD processes recordings of sustained vowels. However, a sustained vowel does not reflect the patients’ everyday communication. This is why the analysis of spontaneous 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 [13, 14]. Technically this is achieved by a ‘prosody module’. The prosody module computes features based on frequency, duration and speech energy measures. This is ‘state-of-the-art’ in automatic speech analysis of normal laryngeal voices [15–17].

In this article we present the results of an automatic evaluation of TE voices with special emphasis on the comparison of data derived from sustained vowels versus text-based analysis. The main question was whether text-based evaluation can replace the vowel-based method.

### Material and Methods

#### Subjects

Audio files were recorded from 24 male laryngectomees (age 60.6 ± 8.9 years) with TE substitute voice. They had undergone total laryngectomy because of laryngeal or hypopharyngeal cancer and had been provided with a Provox shunt valve. At the time of investigation, none of the test persons suffered from recurrent tumor growth or metastases.

#### Hoarseness Diagram

Five sustained vowels (/a/, /e/, /i/, /o/, /u/) were recorded from each patient with a commercial microphone provided with the HD. The program computes parameters called irregularity (irreg), fundamental frequency (F₀), jitter, shimmer, glottal-to-noise excitation ratio (GNE) [18], mean correlation of two adjacent waveform periods (‘p-corr’ in table 1) and noise, which are combined to form the two coordinates in the graphical output. These measures were computed for all five recorded vowels.

#### Text Samples

Each person read the text ‘Der Nordwind und die Sonne’ (‘The North Wind and the Sun’) [19], a standard text consisting of 108 words (71 disjunctive) and 172 syllables used in clinical speech evaluation in German-speaking countries. Data were recorded with a close-talking microphone (dnt Call 4U Comfort headset; DNT GmbH, Dietzenbach, Germany) and digitized with 16 bit at

### Table 1. Correlation between the measures obtained by the HD on recordings of different vowels and prosodic features obtained by the prosody module on the entire text

<table>
<thead>
<tr>
<th>Avg.</th>
<th>a</th>
<th>e</th>
<th>i</th>
<th>o</th>
<th>u</th>
<th>HD</th>
<th>Prosodic feature</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>irreg</td>
<td>maximum length of voiced section</td>
</tr>
<tr>
<td>-0.82</td>
<td>-0.63</td>
<td>-0.67</td>
<td>-0.59</td>
<td>-0.84</td>
<td>-0.69</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-0.79</td>
<td>-0.58</td>
<td>-0.63</td>
<td>-0.60</td>
<td>-0.79</td>
<td>-0.68</td>
<td>irreg</td>
<td>total length of voiced sections</td>
</tr>
<tr>
<td>-0.79</td>
<td>-0.59</td>
<td>-0.64</td>
<td>-0.59</td>
<td>-0.74</td>
<td>-0.70</td>
<td>irreg</td>
<td>ratio length voiced/unvoiced sections</td>
</tr>
<tr>
<td>-0.77</td>
<td>-0.66</td>
<td>-0.61</td>
<td>-0.61</td>
<td>-0.74</td>
<td>-0.60</td>
<td>irreg</td>
<td>voice offset position in current word</td>
</tr>
<tr>
<td>-0.74</td>
<td>-0.59</td>
<td>-0.56</td>
<td>-0.64</td>
<td>-0.67</td>
<td>-0.63</td>
<td>irreg</td>
<td>ratio length voiced sections/length of recording</td>
</tr>
<tr>
<td>-0.81</td>
<td>-0.69</td>
<td>-0.60</td>
<td>-0.63</td>
<td>-0.74</td>
<td>-0.69</td>
<td>jitter</td>
<td>voice offset position in current word</td>
</tr>
<tr>
<td>-0.79</td>
<td>-0.68</td>
<td>-0.64</td>
<td>-0.57</td>
<td>-0.75</td>
<td>-0.67</td>
<td>jitter</td>
<td>maximum length of voiced section</td>
</tr>
<tr>
<td>-0.76</td>
<td>-0.65</td>
<td>-0.61</td>
<td>-0.58</td>
<td>-0.65</td>
<td>-0.68</td>
<td>jitter</td>
<td>ratio length voiced/unvoiced sections</td>
</tr>
<tr>
<td>-0.76</td>
<td>-0.62</td>
<td>-0.61</td>
<td>-0.56</td>
<td>-0.69</td>
<td>-0.66</td>
<td>jitter</td>
<td>total length of voiced sections</td>
</tr>
<tr>
<td>-0.73</td>
<td>-0.60</td>
<td>-0.49</td>
<td>-0.70</td>
<td>-0.65</td>
<td>-0.59</td>
<td>jitter</td>
<td>ratio length voiced sections/length of recording</td>
</tr>
<tr>
<td>-0.79</td>
<td>-0.78</td>
<td>-0.63</td>
<td>-0.22</td>
<td>-0.63</td>
<td>-0.48</td>
<td>F₀</td>
<td>maximum length of voiced section</td>
</tr>
<tr>
<td>-0.77</td>
<td>-0.71</td>
<td>-0.59</td>
<td>-0.24</td>
<td>-0.63</td>
<td>-0.54</td>
<td>F₀</td>
<td>ratio length voiced/unvoiced sections</td>
</tr>
<tr>
<td>-0.76</td>
<td>-0.72</td>
<td>-0.58</td>
<td>-0.24</td>
<td>-0.63</td>
<td>-0.50</td>
<td>F₀</td>
<td>total length of voiced sections</td>
</tr>
<tr>
<td>-0.73</td>
<td>-0.60</td>
<td>-0.47</td>
<td>-0.35</td>
<td>-0.58</td>
<td>-0.60</td>
<td>F₀</td>
<td>voice offset position in current word</td>
</tr>
<tr>
<td>-0.81</td>
<td>-0.78</td>
<td>-0.70</td>
<td>-0.54</td>
<td>-0.77</td>
<td>-0.66</td>
<td>shimmer</td>
<td>maximum length of voiced section</td>
</tr>
<tr>
<td>-0.78</td>
<td>-0.71</td>
<td>-0.66</td>
<td>-0.58</td>
<td>-0.71</td>
<td>-0.68</td>
<td>shimmer</td>
<td>ratio length voiced/unvoiced sections</td>
</tr>
<tr>
<td>-0.76</td>
<td>-0.71</td>
<td>-0.64</td>
<td>-0.54</td>
<td>-0.73</td>
<td>-0.64</td>
<td>shimmer</td>
<td>total length of voiced sections</td>
</tr>
<tr>
<td>-0.75</td>
<td>-0.68</td>
<td>-0.63</td>
<td>-0.59</td>
<td>-0.68</td>
<td>-0.64</td>
<td>shimmer</td>
<td>ratio length voiced sections/length of recording</td>
</tr>
<tr>
<td>+0.70</td>
<td>+0.33</td>
<td>+0.48</td>
<td>+0.64</td>
<td>+0.68</td>
<td>+0.69</td>
<td>p-corr</td>
<td>ratio length voiced/unvoiced sections</td>
</tr>
<tr>
<td>+0.71</td>
<td>+0.35</td>
<td>+0.49</td>
<td>+0.61</td>
<td>+0.74</td>
<td>+0.65</td>
<td>p-corr</td>
<td>maximum length of voiced section</td>
</tr>
</tbody>
</table>

All measure-feature pairs are given where for at least one vowel the correlation was |r| ≥ 0.7. Avg. = Average across all vowels; p-corr = mean correlation of two adjacent waveform periods.
known as mel-frequency cepstrum coefficients are achieved by a discrete cosine transform; these measures are equally spaced on an auditory-based mel scale. The final features are portions in such a section are summed up in intervals (frames) 16 ms in length at a frame shift rate of 10 ms. The frequency cepstrum coefficients, an intensity measure (speech corresponds to the generations of the word hypotheses graphs consisted of the 71 words of the text 'Der Nordwind und die Sonne'. The prosody module derives 95 ‘local’ features for each processed word and 15 ‘global’ features per recording, i.e. on the entire text. 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 on 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 the definition of the local features are given in Bätliner et al. [15, 23] and Zeissler et al. [24].

Results

Table 1 shows the correlations between the frequency-based measures on a sustained vowel calculated by the HD and those obtained in the standard text by the prosody module. 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 correlate with the prosodic features only with $|r| \leq 0.5$ and were therefore not examined further.

Figure 1 compares the average F0 calculations of the HD with those of the prosody module among all recorded sustained vowels of the respective speaker. The prosody module could not compute the frequency for all of the recordings (vowel 'a': n = 6; 'e' and 'i': n = 1; 'u': n = 2 cases) while the HD always gave a value.

Discussion

F0 Computation

For the calculation of jitter and ‘irregularity’ in general F0 plays an important role. In recordings of substitute voice, it is usually difficult to extract periodic sections as previously reported by Moerman et al. [25]. Due to the high irregularity in the vowel recordings, the F0 detection algorithm of the prosody module failed for some of the speakers. The HD computed an F0 value for all recordings, however, sometimes outside a valid range: frequencies of 200 or even 400 Hz are very likely caused by octave errors due to the high irregularity of TE voices. In the study by Olthoff et al. [10] where also different vowels were recorded, reasonable values for F0 could not be computed either. Hence, at present there is no valid method for F0 computation in TE voices.
Correlation between Vowel-Based and Text-Based Measures

Usually, pathologic voice quality is evaluated automatically on recordings of sustained vowels only. For hoarseness, however, it has already been confirmed that acoustic parameters from connected speech are more reliable than those from sustained vowels [26]. For TE speech, it has been shown that automatic analysis of prosodic features of a read out text shows strong correlation with human evaluation criteria, like ‘intelligibility’, ‘speaking effort’ or ‘match of breath and sense units’ [14]. In this article, we examined whether prosodic features obtained from a standard text correlate with the measures that are computed by the vowel-based HD in order to reveal whether HD analysis can be replaced in clinical practice.

The parameters 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 frequency-based 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 $F_0$ detection algorithm – also the values for $F_0$. When comparing these $F_0$ values of the HD with the text-based $F_0$-based prosodic features, no correlations of $|r| \geq 0.7$ were found. Obviously this is due to the fact that the $F_0$ extraction algorithm of the HD does not work adequately with highly pathologic voices and yields invalid values for very low quality voices while the prosody module gives 0, i.e. it decides that the respective section is ‘unvoiced’.

For almost all features, the best correlation is achieved when the average value of the respective HD measure on all 5 vowels is compared with the specific prosodic feature. But although most of the HD measures are highly correlated with prosodic features, no vowel reached $|r| \geq 0.5$ for HD measures GNE and noise which represent the two axes in the graphical output of the program. These measures are a combination of single HD measures [18]. The combination of prosodic features to match GNE and noise better has not been examined yet because the number of possible combinations is tremendous when 95 different local and 15 global features are involved.

It has been shown that the human evaluation criterion ‘match of breath and sense units’ correlates closely with several duration and pause features of the prosody module [14]. There is no sufficient correlation between HD features and pause features of the prosody module. This means that the individual speaking properties 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 single-vowel recordings. The examined prosodic features cannot replace the HD measures completely because the prosody module computes averaged features for a text that contains many different vowels and consonants. On the other hand, the prosody module is also able to perform speech analysis, which means that important aspects like the rhythm of breathing and articulation rate can be estimated automatically. Furthermore, it does not require special hardware. 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 adequately.

The only local prosodic feature that shows high correlation with the HD measures so far is voice offset position in the currently processed word. It denotes the duration between the voice offset and the end of the word, i.e. the length of the unvoiced section at the end of the word. Future work will focus on the local features in order to determine the position of single vowels and compute the frequency-based measures from these vowels in the text. The detected values will also be compared to subjective ratings from human listeners, and patients with substitute voice will be compared with persons with laryngeal hoarseness and normal speakers. The experiments will also involve recordings in languages other than German since the speech recognition system and the prosody module can also be applied to other texts than ‘Der Nordwind und die Sonne’ in other languages.

References

26 Halberstam B: Acoustic and perceptual parameters relating to connected speech are more reliable measures of hoarseness than parameters relating to sustained vowels. ORL 2004;66:70–73.