CHAPTER 5

Acoustic measures and signal typing of voice quality in tracheoesophageal speech, and their relations to perceptual evaluations*

Corina J. van As¹,²
Florien J. Koopmans-van Beinum²
Frans J.M. Hilgers¹
Louis C.W. Pols⁵

¹ Department of Otolaryngology-Head&Neck Surgery, The Netherlands Cancer Institute/ Antoni van Leeuwenhoek Hospital, Amsterdam
² Institute of Phonetic Sciences, University of Amsterdam

*An adapted version of this chapter is submitted for publication
ABSTRACT

In this chapter the results of acoustic analyses (periodicity, harmonicity, and spectral analyses) of tracheoesophageal voice quality are described. First, the results of two software programs (MDVP and Praat) are compared. It appeared that with the program Praat more appropriate calculations could be performed for a larger number of patients, this program is therefore preferred and used for further analysis. Analyses were performed on the sustained vowel /a/. The acoustic measures used are median fundamental frequency, standard deviation of fundamental frequency, jitter, percentage unvoiced, harmonics-to-noise ratio, glottal-to-noise excitation ratio and band energy difference. Next to the use of these acoustic measures, also the use of narrow-band spectrograms for acoustic signal typing of the tracheoesophageal voices into four different types (I-IV) is proposed. Another objective measure used, was the maximum phonation time of sustained /a/. The results of the acoustic signal typing, the acoustic measures, and maximum phonation time are related to the results of the perceptual evaluations reported in Chapter 4 and to the sociodemographic and clinical factors reported in Chapter 3.

Results show that the software program Praat shows the best pitch extraction results for tracheoesophageal voices. Fundamental frequency calculations were possible in 77% of the voice samples. Four other acoustic measures concerning voice quality could be calculated for the entire patient group. Moderate to strong correlations were found between the perceptual evaluations and the acoustic measures.

It can be concluded that reliable and valid acoustic measures can be obtained that are suitable for objective evaluation of tracheoesophageal voice quality, in relation to the perceptual evaluations.
5.1 INTRODUCTION
In Chapter 4, the results of perceptual evaluations of tracheoesophageal speech were described. Despite the fact that perceptual evaluations are considered to be the 'gold standard' for voice quality evaluation, the disadvantages (being subjective and time consuming) should be mentioned as well. Objective acoustic analyses of voice quality, such as periodicity and spectral analyses might solve this problem. In a pilot study preceding this thesis Van As et al. (1998b) found moderate to strong correlations between the perceptual evaluations of sustained /a/ and the acoustic measures that were computed on the same sustained /a/ with the Multi Dimensional Voice Program (MDVP) (Kay Elemetrics, Lincoln Park, NJ, USA). This pilot study also showed that with MDVP in 30% of the voice samples only very short parts were analyzable due to aperiodicity of the voice. This indicates that for the poorer voices only the short parts that contain some periodicity can be analyzed. In another study by Van As et al. (1998a), investigating the influence of stoma occlusion on tracheoesophageal voice, three voice samples of 1 second of a sustained /a/ were analyzed by MDVP. In 19% of the cases, none of the three voice samples could be analyzed, for another 19% only one or two out of the three voice samples could be analyzed. Visual inspection of the voice samples that could not be analyzed showed that the patients had very low-pitched voices (and thus fell outside the fixed pitch analysis range of MDVP) or had very aperiodic voices. Tracheoesophageal voice can thus be aperiodic to such an extent or can have such an extremely low pitch, that the pitch detection algorithm fails, or that there is no fundamental frequency present at all. This implies that when acoustic measures are used that are based on pitch detection algorithms, reliable results can be obtained only for the better tracheoesophageal voices.

The aim of the acoustic analyses in the present study is thus first of all to determine whether acoustic measures can be found that allow calculations for the entire range of tracheoesophageal voice qualities. If, subsequently, these measures can be reliably obtained for the entire range of voice qualities of the patients in the study, they can be related to the results of the perceptual evaluations. Correlation coefficients will then show whether these measures can be considered suitable for objective evaluation of voice quality.

For this purpose, in view of the findings in the pilot study (Van As et al. 1998b) and in view of the results in the study on stoma occlusion (Van As et al. 1998a), not only the Computerized Speech Lab with the MDVP program was used, but, additionally, the program Praat, developed at the Institute of Phonetic Sciences, University of Amsterdam (Boersma and Weenink 1996).

5.2 PATIENTS AND METHODS

5.2.1 PATIENTS
Information about the speakers participating in this part of the study is documented in Chapter 3. Briefly, the speakers were 39 laryngectomized speakers: 29 males and 10 females. See also Table 3.1.

5.2.2 SPEECH MATERIAL/RECORDING/PROCESSING
The speech material for acoustic analyses consisted of three sustained vowels /a/ at comfortable pitch and loudness level, and three attempts for maximum phonation time, also on sustained /a/. The speech recordings were made in a quiet, sound-treated room. For the recordings, a DAT-recorder (Sony TCD-8) was used, together with the hardware and software of the Computerized Speech Lab, Model 4300B (Kay Elemetrics, Lincoln Park, NJ, USA). Via the external module of the Computerized Speech Lab the speech data were
digitally recorded on DAT tape. A headset microphone (AKG-c410) was used; the mouth-microphone distance was 2.5 centimetres. At the beginning of the speech recording, the recording level was adjusted for each speaker separately and then fixed, in order to optimize the signal-to-noise ratio.

For analysis, the three sustained /a/ s for acoustic analyses and the three /a/ s for maximum phonation time, were stored on a PC hard disk with a sampling frequency of 44100 Hz using a SoundBlaster card and the Praat software. The two software programs used for periodicity and spectral analyses are the Multi Dimensional Voice Program (MDVP) (Kay Elemetrics Corp., Lincoln Park, NJ, USA) and Praat (Praat, a system for doing phonetics by computer, Version 3.8.68, web site: www.praat.org, Copyright 1992-2000 by P. Boersma and D. Weenink). The MDVP software is often used as a clinical tool for objective voice evaluation and calculates 32 different voice quality measures. The program Praat is a research, publication, and productivity tool for phoneticians.

5.2.3 ACOUSTIC SIGNAL TYPING

Using Praat, a narrow-band spectrogram (100 ms window) was made of each of the three sustained /a/ s of each speaker. Before performing further analyses the sustained /a/ voice samples of the speakers were divided into subgroups based on the visual appearance of the acoustic signal in a narrow-band spectrogram. Titze (1994) recommends such a signal typing and draws conclusions from the type of signal for possibilities of further acoustic analysis like perturbation analysis. Acoustic signal typing according to Titze is as follows: Type-I signals are nearly-periodic; Type-II signals contain intermittence, strong sub-harmonics or modulations; and Type-III signals are chaotic or random. Perturbation measures can be performed reliably only in the Type-I voice signals.

Visual inspection of the narrow band spectrograms of the sustained /a/ s of the tracheoesophageal speakers in the present study showed that one of the more obvious differences among the speakers lies in the ‘harmonic strength’ of the signal. In some speakers harmonics up to 1000 Hz were seen, while in other speakers only one or two harmonics, or even no harmonics at all, were seen. The amount of spectral noise thus differs among the patients. Therefore, based on the spectrographic characteristics of the tracheoesophageal voice signals, we defined our own criteria for these tracheoesophageal voices. The criteria used for the acoustic signal typing were as follows:

**Type I. Stable & Harmonic** (see Figure 5.2)
- Stable signal for longer than two seconds, and
- Clear harmonics up to at least 1000 Hz

**Type II. Stable & At least one harmonic** (see Figure 5.3)
- Stable signal for longer than two seconds, and
- At least one stable harmonic at the fundamental frequency for longer than two seconds

**Type III. Unstable or Partly Harmonic** (see Figure 5.4 and Figure 5.5)
- No stable signal for longer than two seconds, or
- Harmonics in only part of the sample (for longer than 1 second)

**Type IV. Barely Harmonic** (see Figure 5.6)
- No or only short-term detectable harmonics (for shorter than 1 second)
In Figure 5.1, a schematic graphic representation of the acoustic signal types is given. The type-I signals, being the voice samples with the best harmonic structure and a stable fundamental frequency, are situated in the upper right corner of the diagram at the positive side of the x-axis as well as at the positive side of the y-axis. The type-II signals, showing only one or a few harmonics with a stable fundamental frequency, are placed at the middle of the x-axis (not showing harmonics up to at least 1000 Hz, but also not entirely consisting of noise) and at the positive side of the y-axis, having a stable fundamental frequency. The type-III signals are always at the negative side of the y-axis, being unstable, and can present at a large part of the x-axis when the fundamental frequency shows to be unstable or when the voice sample shows both harmonic and noisy parts. The type-IV signals always consist for the largest part of noise and are barely harmonic, thus they are represented at the negative end of the x-axis, and since they do not contain a fundamental frequency; the stability of the fundamental frequency (y-axis) is not an issue in these signal types.

Figure 5.1. Schematic graphic representation of the four acoustic signal types based on the harmonic structure and stability of the signal.
Figure 5.2. Example of a voice sample classified as type I. The oscillogram shows a stable signal, with stable loudness. The 100 ms selection of the oscillogram shows a clearly periodic pattern, with some noise. The pitch contour shows a stable fundamental frequency (mean 122 Hz). In the narrow-band spectrogram (100 ms window) clear harmonics are seen up to 1500 Hz and for parts of the voice sample even up to 2000 Hz. The long-term average spectrum over 1 s also shows a clear harmonic structure in the lower frequencies and noise in the higher frequency region.
Figure 5.3. Example of a voice sample classified as type II. The oscillogram shows a stable signal, with stable loudness. The 100 ms selection of the oscillogram shows a periodic pattern, with noise. The pitch contour shows a stable fundamental frequency (78 Hz). In the narrow-band spectrogram (100 ms window) the first harmonic is clearly visible, the second and third harmonic are visible in small parts of the spectrogram. In the long-term average spectrum over 1 s also only 3 harmonics can be seen, the high-frequency noise is of a higher level than in the type-I signal in Figure 5.2.
Figure 5.4. Example of a voice sample classified as type III. The oscillogram shows some signs of instability in the signal. According to the pitch contour a fundamental frequency is only found in a part of the signal. The narrow-band spectrogram (100 ms window) shows that the first harmonic is only clearly present in the first part of the voice sample. In the long-term average spectrum over 1 s, four to five harmonics can be seen, the level of the high-frequency noise is high.
Figure 5.5. Second example of a voice sample classified as type III. The oscillogram shows a signal of unstable loudness. The 100 ms selection of the oscillogram shows a clearly periodic structure. The pitch contour shows that the patient is unable to produce a sustained /a/ at a stable pitch. The narrow-band spectrogram (100 ms window) shows that there are clear harmonics up to 2000 Hz, but that the voice signal is very unstable. In the long-term average spectrum over 1 s only four harmonics are seen, and due to the instability, the noise in the spectrum is high.
Figure 5.6. Example of a voice sample classified as type IV. The oscillogram shows a highly unstable signal. In the 100 ms selection of the oscillogram, no periodicity can be detected at all. This is reflected in the pitch contour, the narrow-band spectrogram (100 ms window) and the long-term average spectrum over 1 s in which no harmonicity is seen at all.
For each patient the vowel recording with the best signal type was selected from the three available vowels for the acoustic signal typing and for further acoustic analysis.

### 5.2.4 Acoustic analyses

First, of the vowel that was selected for further analysis, two seconds of the most stable part, as seen in de narrow band spectrogram, were selected for analysis.

The software program Praat was used to calculate:

- **The median fundamental frequency** (F0-median (Hz)).
  The fundamental frequency as a function of time was measured using cross-correlation. Default settings were used, except for the pitch extraction range, which was set from 40 Hz to 250 Hz, and the voicing threshold, which was set to 0.40 instead of 0.45. The length of the analysis window is based on the value of the lowest frequency of the pitch extraction range. In order to avoid pitch extraction errors, this frequency was increased for higher pitched voices. The slight decrease of the voicing threshold enables a ‘voiced’ decision in a larger part of the voice sample, a larger decrease of this voicing threshold was found to introduce pitch extraction errors as observed by visual inspection. The median fundamental frequency is determined over the voiced segments of the 2-second interval.

- **The standard deviation of the fundamental frequency** (F0-SD (Hz)).
  The standard deviation of the fundamental frequency is derived from the fundamental frequency measurement and reflects the changes in fundamental frequency found in the two-second voice sample of sustained /a/.

- **Jitter** (%).
  The percentage of jitter was calculated from the results of the pitch extraction. First, a point process was created from the results of the pitch extraction. Then, the pulses extracted from the point process were used to calculate jitter. The shortest possible interval that was considered was 0.1 ms. The longest possible interval considered was related to the lowest fundamental frequency found for that particular voice sample (for example when the lowest F0 found was 50 Hz, this period was set to 20 ms).

- **The percentage of voiced** (%Voiced).
  This percentage was calculated on the relative number of unvoiced analysis windows found for the calculation of the fundamental frequency. In voice samples where no fundamental frequency was found, this percentage was thus zero.

- **The harmonics-to-noise ratio** (HNR (dB)).
  The HNR was calculated using cross-correlation. Default settings were used, except for minimum frequency for pitch extraction, which was set to 40 Hz, and for the silence threshold, which was set to zero. With the silence threshold set to zero, the harmonics-to-noise ratio calculated is based on the entire voice sample and not only on the louder parts.

Then, a stable part of 0.25 s with the best harmonic structure was selected from the 2-second voice sample and the program Praat was used to calculate:

- **The glottal-to-noise excitation ratio** (GNE).
  This measure indicates to what extent the voice excitation is due to a pulse train or due to noise (Michaelis et al., 1997). The minimum frequency was 500 Hz, the maximum frequency 4500 Hz, the frequency band used was 1000 Hz, and the frequency step was 80 Hz.
The 0.25 s part was then sampled down to 10 kHz and a long-term average spectrum was made and used to calculate:

The **band energy difference** (BED (dB)).

This is the difference in decibels between the mean spectral intensity in the band between 0 to 500 Hz and the mean spectral intensity in the band between 4000 and 5000 Hz in the spectrum; it can be considered as an estimation of the relative amount of high-frequency noise in the spectrum of the voice (Debruyne et al., 1994; Dejonckere and Lebacq, 1987).

The software program MDVP Advanced (Multi Dimensional Voice Program) of Kay Elemetrics Corp. was also used for acoustic analysis of the 2-second voice samples. With this program 32 different voice measures are calculated, mainly covering fundamental frequency, frequency and amplitude, voice breaks, irregularities, subharmonics, noise, and tremor aspects. In Van As et al. (1998b) all parameters measured, are described in detail. Since it was decided (see section 5.3.2.1) not to use MDVP for the acoustic analyses, no further details are given here.

**5.2.5 Maximum phonation time**

Also, the maximum phonation time was measured. The longest possible phonation of the three attempts was used for further evaluation. The length in seconds of the longest /a/ was measured with Praat and rounded to the nearest integer.

**5.2.6 Statistical analyses**

Statistical analyses were performed using the Statistical Package for Social Sciences, version 10.0 (SPSS Inc., Chicago, Ill.). A p-value below .05 is considered significant. Means and standard deviations of the acoustic measures were calculated for the entire patient group. Before further statistical analyses were performed, the acoustic measures concerning fundamental frequency (median fundamental frequency and standard deviation of fundamental frequency) were logarithmically transformed in order to be more perceptually relevant. The numbers in text and Tables are transformed back to Hz for clearness sake.

Analyses of variance (ANOVA) followed by post hoc Tukey tests were performed to investigate the relations between the judgment of overall voice quality and the acoustic measures obtained with Praat, and to investigate the relations between the acoustic signal types and the acoustic measures. When assumptions of normality could not be met (according to Q-Q plots) the non-parametric Kruskal-Wallis test followed by the non-parametric Mann-Whitney test with Bonferoni correction was used.

The relation between the acoustic signal types and the overall perceptual judgment of tracheoesophageal voice quality by the trained raters (good-reasonable-poor) was investigated with a chi-squared test for linear-by-linear association.

Pearson’s correlation coefficients were computed to investigate the relations between the acoustic measures obtained with Praat and maximum phonation time, and relations between the acoustic measures and the perceptual scale judgments (for perceptual evaluations, see Chapter 4). A correlation coefficient between .10 and .30 is considered weak, a coefficient between .30 and .50 moderate, and a coefficient higher than .50 strong (Meuffels, 1992).

Relations between the clinical factors (reconstruction, age, postoperative follow-up, myotomy, neurectomy, radiotherapy, and radical neck dissection) and the acoustic signal typing were investigated by chi-squared tests.

Relations between the clinical factors and the acoustic measures obtained with Praat and relations between the clinical factors and maximum phonation time were investigated with a t-test for independent samples, or, when the assumption of normality cannot be met
Acoustic analyses versus perceptual evaluations

(according to Q-Q plots) with the Man-Whitney U test for non-parametric measures. More information on the clinical factors and the subgroups that are based on them is given in Chapter 3.

5.3 RESULTS

First, the results of the three separate parts of the acoustic analyses will be described in the following order:

- acoustic signal typing (section 5.3.1);
- acoustic measures (MDVP and Praat), including a comparison between MDVP and Praat (section 5.3.2);
- maximum phonation time (section 5.3.3).

After comparison of the results of MDVP and Praat, Praat was chosen for the acoustic measures. Thus, from this point on, when talking about acoustic measures, those obtained with Praat are meant.

Relations between the four acoustic signal types and the acoustic measures and maximum phonation time will be described in section 5.3.4.

After describing the acoustic analyses and the relations amongst them, relations with the perceptual evaluations (overall judgment of voice quality as good, reasonable, or poor and judgment of 19 semantic bipolar 7-point scales, see Chapter 4) are investigated. These relations are described in the following order:

- acoustic signal typing versus perceptual evaluations (section 5.3.5);
- acoustic measures versus perceptual evaluations (section 5.3.6);
- maximum phonation time versus perceptual evaluations (section 5.3.7).

In section 5.3.8 the three parts of the acoustic analyses (acoustic signal typing, acoustic measures, and maximum phonation time) will be related to the sociodemographic and clinical factors (see Chapter 3).

5.3.1 ACOUSTIC SIGNAL TYPING

On the basis of the criteria used for the acoustic signal typing, four subgroups were formed. In Table 5.1 the number of patients for each of the signal types is shown.

Table 5.1. Acoustic signal typing of the voice samples of the sustained /a/ for all 39 speakers.

<table>
<thead>
<tr>
<th>Type</th>
<th>Definition</th>
<th>Number of patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>stable &amp; harmonic</td>
<td>7</td>
</tr>
<tr>
<td>II</td>
<td>stable &amp; at least one harmonic</td>
<td>13</td>
</tr>
<tr>
<td>III</td>
<td>unstable or partly harmonic</td>
<td>11</td>
</tr>
<tr>
<td>IV</td>
<td>barely harmonic</td>
<td>8</td>
</tr>
</tbody>
</table>

5.3.2 ACOUSTIC MEASURES

5.3.2.1 MDVP versus Praat

Results for MDVP and Praat differed. In MDVP 16 out of the 39 two-second voice samples of sustained /a/ were considered unvoiced. Visual inspection learned that this was not always correct. Actually, in Praat only 6 out of these 16 voice samples (that were considered unvoiced in MDVP) were considered unvoiced. The main reason for this difference is the fixed pitch extraction range of MDVP (70 Hz to 625 Hz). Praat is more suitable in that
CHAPTER 5

respect, since the pitch extraction range can be chosen by the investigator. In Table 5.2 an overview is given of the differences between the pitch extraction results of MDVP and Praat.

Table 5.2. Overview of the differences between the pitch extraction results of MDVP and Praat.

<table>
<thead>
<tr>
<th>MDVP-Pitch Extraction</th>
<th>Praat-Pitch Extraction</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unvoiced: n=16</td>
<td>Voiced (n=10)</td>
<td>F0 (Hz) = 45, 50, 52, 64, 66, 66, 75, 78, 81, 227.</td>
</tr>
<tr>
<td></td>
<td>Unvoiced (n=6)</td>
<td></td>
</tr>
<tr>
<td>Voiced: n=16</td>
<td>Voiced: n=16</td>
<td>Equal F0: n=6 difference in F0 (Hz) = 1, 1, 1, 2, 2, 4, 4, 5, 5, 15.</td>
</tr>
<tr>
<td>Pitch Extraction errors: n=7</td>
<td>Voiced: n=4</td>
<td>F0 (Hz) = 76, 159, 163, 174.</td>
</tr>
<tr>
<td></td>
<td>Unvoiced: n=3</td>
<td></td>
</tr>
</tbody>
</table>

In Table 5.2 it can be seen that 10 out of the 16 voice samples, which were considered unvoiced within MDVP, were considered voiced within Praat. Of 9 of these voice samples, the fundamental frequency was most probably too low for analysis with MDVP. They were considered unvoiced with MDVP, but the mean fundamental frequencies calculated with Praat for the voice samples of those 9 patients were 45, 50, 52, 64, 66, 66, 75, 78 and 81 Hz. One voice sample, which was considered unvoiced in MDVP was found to have a fundamental frequency of 227 Hz in a short part of the voice sample with Praat. Furthermore, visual inspection of the pitch extraction of the 23 voice samples that could be analyzed with MDVP, showed that 16 voice samples were analyzed correctly and that some pitch extraction errors occurred in 7 voice samples. In one sample this was due to an octave jump, in the remaining 6 samples calculations were based on sometimes very short parts of the sample (0.04 to 0.55 seconds) in which the pitch extraction was incorrect. Analysis of these 7 voice samples in Praat, taught us that of these 7 samples, 3 were actually considered unvoiced within Praat, one was analyzed with a very low frequency (76 Hz), and three were analyzed correctly with fundamental frequencies of 159, 163, and 174 Hz. In 16 patients results could be computed with both MDVP and Praat. Comparison of the results for the mean fundamental frequency showed that the results are comparable. In 6 voice samples results of both programs were equal, in 3 there was a difference of 1 Hz, in 2 a difference of 2 Hz, in 2 a difference of 4 Hz, in 2 a difference of 5 Hz, and in 1 a difference of 15 Hz. The latter voice sample was very unstable: with MDVP some very high frequencies were measured that were not measured with Praat.

Overall, using Praat, 9 voice samples out of 39 (23%) can be considered unvoiced, visual inspection of these 9 voice samples indeed shows no clear periodicity. Using MDVP, a total of 23 voice samples out of 39 (59%) are considered unvoiced (16 by the program itself and 7 by the investigator due to pitch extraction errors). These results and the fact that with MDVP no adjustments can be made by the user to improve the results for the 32 measures, led to the decision to use Praat for the further evaluations. It should however be mentioned in this respect that the adjustments that can be made in Praat to optimize the results, might also lead to inconsistencies when used incorrectly.

5.3.2.2 Acoustic analyses with Praat

In Table 5.3, the global results for the acoustic measures computed with Praat are given. The acoustic measures based on pitch extraction could be computed for the 30 voice samples of
Acoustic analyses versus perceptual evaluations

the patients that were considered “voiced”; these measures are the median fundamental frequency (F0-median), the standard deviation of the fundamental frequency (F0-SD), and jitter. The acoustic measures not based on pitch extraction are computed for all 39 voice samples; these measures are the percentage of voiced (%Voiced), the harmonics-to-noise ratio (HNR), the glottal-to-noise excitation ratio (GNE), and the band energy difference (BED). Although the percentage of voiced is based on the results of pitch extraction, it could be computed for the entire patient sample since it was considered zero when no pitch was measured.

Table 5.3. Means, range and standard deviations of the acoustic measures. The entire patient group consists of 39 (29 males, 10 females) patients. Median fundamental frequency (F0-median), the standard deviation of the fundamental frequency (F0-SD), and Jitter could be calculated for voice samples of 30 patients (9 voice samples of patients were considered unvoiced). The remaining measures, percentage of voiced (%Voiced), harmonics-to-noise ratio (HNR), glottal-to-noise excitation ratio (GNE), and band energy difference (BED) could be calculated for all voice samples of all 39 patients.

<table>
<thead>
<tr>
<th>Acoustic parameter</th>
<th>N</th>
<th>Range</th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>F0-median (Hz)</td>
<td>30</td>
<td>46 to 229</td>
<td>103</td>
<td>43</td>
</tr>
<tr>
<td>F0-SD (Hz)</td>
<td>30</td>
<td>0.09 to 35.8</td>
<td>6.31</td>
<td>7.17</td>
</tr>
<tr>
<td>Jitter (%)</td>
<td>30</td>
<td>0 to 100</td>
<td>66</td>
<td>40</td>
</tr>
<tr>
<td>%Voiced (%)</td>
<td>39</td>
<td>0 to 100</td>
<td>4.30</td>
<td>4.45</td>
</tr>
<tr>
<td>HNR (dB)</td>
<td>39</td>
<td>2.0 to 15.3</td>
<td>7.8</td>
<td>11</td>
</tr>
<tr>
<td>GNE</td>
<td>39</td>
<td>0.57 to 98</td>
<td>-19.0</td>
<td>9.3</td>
</tr>
<tr>
<td>BED (dB)</td>
<td>39</td>
<td>-37.3 to -2.5</td>
<td>-19.0</td>
<td>9.3</td>
</tr>
</tbody>
</table>

Results show that for all acoustic measures the range is wide and the standard deviation is high. The variability amongst the patients is thus high. The median fundamental frequency varies from 46 to 229 Hz. The median fundamental frequency of the entire group (103 Hz) is comparable to the mean fundamental frequency of normal male speakers (110 Hz) (see section 5.4). In this respect it is noteworthy to refer to the absence of a difference in fundamental frequency between male and female tracheoesophageal speakers. In the present study, even the lowest median fundamental frequency of 46 Hz was produced by a female tracheoesophageal speaker, whereas the highest one of 229 Hz was produced by a male speaker.

5.3.3 Maximum Phonation Time

The maximum phonation time of a sustained /a/ varied from 3 s to 37 s. The mean maximum phonation time was 12.8 s and the standard deviation was 7.9 s.

5.3.4 Acoustic Signal Typing Versus Acoustic Measures and Maximum Phonation Time

Regarding the acoustic measures percentage of voiced, glottal-to-noise excitation ratio, harmonics-to-noise ratio, and band energy difference as well as for maximum phonation time relations could be computed for all speakers. The acoustic measures median fundamental frequency, standard deviation of fundamental frequency, and jitter could be computed for all type-I signals, 12 out of the 13 type-II signals, 10 out of the 11 type-III signals, and only one out of the 8 type-IV signals. The percentage of voiced of this one type-IV voice sample was 18%, indicating that only a short part of this voice sample contained periodicity. This result could be expected since type-IV signals are defined as voice samples with only short-term or no detectable harmonics.
Table 5.4. Results of analyses of variance on the subgroups of acoustic signal typing with the acoustic measures. For the measures percentage of voiced (%Voiced), harmonics-to-noise ratio (HNR), glottal-to-noise-excitation ratio (GNE), band energy difference (BED) and maximum phonation time (MPT) the analyses are based on four subgroups of acoustic signal typing (I (n=7), II (n=13), III (n=11), and IV (n=8)). For the measures F0-median, F0-SD, and jitter the analyses are based on three subgroups of acoustic signal typing (I (n=7), II (n=12), and III (n=10)). The arrows between the values indicate that there is a significant difference between those values; the box attached to the arrow gives the exact p-value.

The type-IV signals are therefore left out of the analysis for these acoustic measures: the group size is too small and the fact that these measures mostly cannot be calculated for this acoustic signal type already indicates the difference between this signal type and the three other types. Since assumptions of normality could not be met for the acoustic measures percentage of voiced and jitter, non-parametric tests were used. The results of the analyses of variance are given in Table 5.4.

From Table 5.4 it becomes clear, that the acoustic measures based on pitch extraction (median fundamental frequency, standard deviation of fundamental frequency and jitter) differentiate type-IV signals from the other signal types simply by the fact that these measures cannot be calculated for the majority of the type-IV voice samples. The standard deviation of the fundamental frequency differentiates between type I and III, being lower in type-I signals. The percentage jitter differentiates between type I and II signals, and between type I and III signals, being lower in type-I signals. None of the acoustic measures was found to differentiate between type-II and type-III signals. The percentage of voiced differentiates between all signal types, except perhaps the types II and III, the harmonics-to-noise ratio separates the type-I signals from the type-II, type-III, and type-IV signals and the type-III signals from the type-IV signals. The band energy difference differs between the type-IV signals and the type-I signals and between the type-IV signals and the type-II signals.
Although the band energy difference is lower in the type-III signals, this difference is not statistically significant. The glottal-to-noise excitation ratio was not found to differentiate between any of the four signal types.

5.3.5 Acoustic signal typing versus perceptual evaluations

In this section relations between the acoustic signal types and the perceptual evaluations are shown. First, in section 5.3.5.1, results are shown for the relations between the classification into the four types based on the narrow band spectrogram and the overall perceptual judgment of tracheoesophageal voice quality as good, reasonable, or poor (see Chapter 4). Then, in section 5.3.5.2 results are shown for the relations between the acoustic signal typing and the perceptual semantic scales as judged by the trained raters.

5.3.5.1 Acoustic signal typing versus overall judgment of voice quality

In Table 5.5 the relation between the acoustic signal typing into the four different types and the overall perceptual judgment of voice quality by the trained expert raters as good, reasonable, or poor is shown. A chi-squared test for linear-by-linear association shows that there is a significant relation (p<.001).

Table 5.5. Table of the relation between the four types of acoustic signal typing and the perceptual judgment of overall voice quality for all 39 speakers. Numbers represent number of patients.

<table>
<thead>
<tr>
<th>Acoustic signal typing</th>
<th>Perceptual judgment of overall voice quality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Good</td>
</tr>
<tr>
<td>Type I</td>
<td>5</td>
</tr>
<tr>
<td>Type II</td>
<td>6</td>
</tr>
<tr>
<td>Type III</td>
<td>2</td>
</tr>
<tr>
<td>Type IV</td>
<td>0</td>
</tr>
</tbody>
</table>

Type-IV signals were never perceived as good, while type-I signals were never perceived as poor. Two patients that show type-III signals were nevertheless perceived as good, they got the qualification type III since their voice sample did not meet the criterion of visible harmonics for longer than two seconds. As can be seen, type-II signals occur both in good and reasonable voices, apparently a voice sample with one stable harmonic and an otherwise noisy spectrum can still be perceived as a good tracheoesophageal voice.

5.3.5.2 Acoustic signal typing versus perceptual scale judgments

One-way analyses of variance (ANOVA) shows that the acoustic signal typing into the four types is significantly related to 14 out of the 17 semantic bipolar 7-point scales that were judged by the trained expert raters (p<.05). The 14 relating scales are: deviant-normal, unpleasant-pleasant, ugly-beautiful, noise-no noise, monotonous-melodious, expressionless-expressive, weak-powerful, unsteady-steady, jerking-fluent, bubbly-not bubbly, breathy-not breathy, hypertonic-not hypertonic, hypotonic-not hypotonic, and unintelligible-intelligible. The scales slow-quick, low-high, and deep-shrill were not found to be related to the acoustic signal types. In Table 5.6 the results of these analyses of variance are given.
Table 5.6. Results of one-way analyses of variance (ANOVA) for the four signal types of acoustic signal typing with the 17 perceptual scale judgments on a semantic bipolar 7-point scale by the trained expert raters (Chapter 4). In column 1 each perceptual scale is represented, in column 2 the p-value of the analysis of variance for that scale, and in column 3-6 the mean value of the scale judgments for the subgroups of acoustic signal typing is given. When with post hoc Tukey tests significant differences are found between the four subgroups, an arrow and the p-value are drawn between the two means of the subgroups. The arrows between the values indicate that there is a significant difference between those values; the box attached to the arrow gives the exact p-value.

<table>
<thead>
<tr>
<th>Perceptual scale</th>
<th>p-value</th>
<th>Acoustic signal typing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>type I (n=7)</td>
</tr>
<tr>
<td>deviant-normal</td>
<td>&lt;.001</td>
<td>5.0</td>
</tr>
<tr>
<td>unpleasant-pleasant</td>
<td>.001</td>
<td>5.2</td>
</tr>
<tr>
<td>ugly-beautiful</td>
<td>.001</td>
<td>5.1</td>
</tr>
<tr>
<td>noise-no noise</td>
<td>.013</td>
<td>5.7</td>
</tr>
<tr>
<td>monotonous-melodious</td>
<td>.001</td>
<td>5.3</td>
</tr>
<tr>
<td>expressionless-expressive</td>
<td>.004</td>
<td>5.9</td>
</tr>
<tr>
<td>weak-powerful</td>
<td>&lt;.001</td>
<td>5.3</td>
</tr>
<tr>
<td>unsteady-steady</td>
<td>.019</td>
<td>5.9</td>
</tr>
<tr>
<td>jerking-fluent</td>
<td>.002</td>
<td>6.3</td>
</tr>
<tr>
<td>slow-quick</td>
<td>.258</td>
<td>4.0</td>
</tr>
<tr>
<td>low-high</td>
<td>.378</td>
<td>4.6</td>
</tr>
<tr>
<td>deep-shrill</td>
<td>.186</td>
<td>4.3</td>
</tr>
</tbody>
</table>
jerking-fluent show lower mean ratings in types III. The mean rating for the scale hypertonic-powerful, bubbly-not bubbly and unintelligible-intelligible.

From Table 5.6 a number of interesting observations can be made. Eight of the perceptual bipolar semantic 7-point scales that show a relation with the acoustic signal types are related to the ranking of the four types. The average ratings of the perceptual scales for the signal types decrease with the deviancy of the signal types for the scales deviant-normal, pleasant-unpleasant, ugly-beautiful, monotonous-melodious, expressionless-expressive, weak-powerful, bubbly-not bubbly and unintelligible-intelligible.

The mean rating for the scale extraneous speaking noise-no extraneous speaking noise (noise-no noise) is lower for type III compared to types I and II. Incomplete stoma occlusion and incorrect timing of the stoma occlusion are known to be the main causes of extraneous speaking noise. The relation found here is thus not surprising since one of the criteria for types III is the absence of a steady signal for longer than two seconds. Understandably, incomplete or incorrectly timed stoma occlusion leads to instability of the voice signal. The inability to produce a steady voice signal for over 2 seconds as seen in the acoustic signal typing also relates to the perceptual impression of instability: the scales unsteady-steady and jerking-fluent show lower mean ratings in types III. The mean rating for the scale hypertonic-not hypertonic is also lower in types III, and thereby the mean rating of the scale hypotonic-not hypotonic and the perceptual scales related to hypotonicity (bubbly-not bubbly, breathy-not breathy) show higher ratings in types III, especially compared to types IV. This indicates that hypertonicity also influences the instability found in types III, while hypotonicity relates to the lack of harmonics in the type-IV signals.

### 5.3.6 Acoustic analyses versus perceptual evaluations

In section 5.3.6.1 relations between the acoustic measures of voice quality calculated with Praat and the overall perceptual judgment of tracheoesophageal voice quality as good, reasonable, or poor are shown. Then, in section 5.3.6.2, results are shown for the relations between the acoustic measures and the separate perceptual scales.
5.3.6.1 Acoustic measures versus overall judgment of voice quality

Relations between the acoustic measures and the perceptual judgment of overall voice quality performed by the trained expert listeners (Chapter 4) were studied by means of analyses of variance. On behalf of assumptions of normality, non-parametric tests were used for the measures band energy difference, percentage of voiced and harmonics-to-noise ratio. In Table 5.7 the results of these analyses are shown.

Table 5.7. Results of the analyses of variance for the three voice quality subgroups and the acoustic measures. For the F0-median, F0-SD and jitter measures the group sizes are: good (n= 12), reasonable (n= 11), poor (n= 7). For the HNR, GNE, BED and % Voiced measures, the group sizes are: good (n= 12), reasonable (n= 14), poor (n= 13). In the second column the overall p-value is given, in column 3-5 means are given for the subgroups. Arrows in between those means indicate the p-level for the difference between the subgroups. The arrows between the values indicate that there is a significant difference between those values; the box attached to the arrow gives the exact p-value.

<table>
<thead>
<tr>
<th>Acoustic parameter</th>
<th>p-value</th>
<th>Voice Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Good</td>
</tr>
<tr>
<td>F0-median (Hz)</td>
<td>.322</td>
<td>91</td>
</tr>
<tr>
<td>F0-SD (Hz)</td>
<td>.025</td>
<td>3.30</td>
</tr>
<tr>
<td>Jitter (%)</td>
<td>.456</td>
<td>6.67</td>
</tr>
<tr>
<td>%Voiced*</td>
<td>.004</td>
<td>89</td>
</tr>
<tr>
<td>HNR*</td>
<td>.024</td>
<td>6.3</td>
</tr>
<tr>
<td>GNE</td>
<td>.461</td>
<td>.80</td>
</tr>
<tr>
<td>BED (dB)*</td>
<td>.008</td>
<td>-24.04</td>
</tr>
</tbody>
</table>

*Non-parametric measures

The standard deviation of the fundamental frequency differs between the good and the reasonable group (p=.020). The band energy difference also differs between the good and the reasonable group (p=.015) and between the good and the poor group (p=.010). The percentage of voiced differs between the good and the poor group (p=.002) and between the reasonable and the good group (p=.022).

5.3.6.2 Acoustic measures versus perceptual scale judgments

Relations between the acoustic measures and the perceptual scale judgments performed by the trained expert listeners (Chapter 4) are studied by means of Pearson’s correlation coefficients. The acoustic measures F0-median, F0-low, F0-high, F0-SD, and jitter could not be calculated for all patients, thus these correlations are based on 30 patients, for the other measures they are based on the entire patient group of 39 patients. In Table 5.8 significant correlations are given. The percentage of jitter was logarithmically transformed before Pearson’s correlation coefficients were calculated.
Table 5.8. Correlation matrix of the acoustic measures (columns) and the perceptual scale judgments (rows). Correlations that are significant below the .05 level are flagged with *, correlations that are significant below the .001 level are flagged with **. For the measures F0-median, F0-SD, and jitter the correlations are based on 30 observations, for the remaining measures on 39 observations.

<table>
<thead>
<tr>
<th>Perceptual Scale</th>
<th>F0-median</th>
<th>F0-SD</th>
<th>Jitter (log)</th>
<th>HNR</th>
<th>%Voiced</th>
<th>GNE</th>
<th>BED</th>
</tr>
</thead>
<tbody>
<tr>
<td>deviant-normal</td>
<td>-.46**</td>
<td></td>
<td>.50**</td>
<td>.57**</td>
<td></td>
<td>-.58**</td>
<td></td>
</tr>
<tr>
<td>unpleasant-pleasant</td>
<td>-.53**</td>
<td>.38**</td>
<td>.42**</td>
<td></td>
<td></td>
<td>-.57**</td>
<td></td>
</tr>
<tr>
<td>ugly-beautiful</td>
<td>-.46**</td>
<td>.43**</td>
<td>.50**</td>
<td></td>
<td></td>
<td>-.64**</td>
<td></td>
</tr>
<tr>
<td>noise-no noise</td>
<td>-.51**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>monotonous-melodious</td>
<td>-.45**</td>
<td>.38**</td>
<td>.40**</td>
<td>-.53**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>expressionless-expressive</td>
<td>-.45**</td>
<td>.55**</td>
<td>.68**</td>
<td>.40**</td>
<td>-.55**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>weak-powerful</td>
<td>-.46**</td>
<td>.54**</td>
<td>.67**</td>
<td>-.53**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>unsteady-steady</td>
<td>-.52**</td>
<td>.54**</td>
<td>.67**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>jerking-fluent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>slow-quick</td>
<td>.63**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>low-high</td>
<td>.54**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>deep-shrill</td>
<td>-.38*</td>
<td>.48**</td>
<td>.57**</td>
<td>.33**</td>
<td>-.38**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bubbly-not bubbly</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>breathy-not breathy</td>
<td>.49**</td>
<td>.54**</td>
<td>.67**</td>
<td>-.54**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hypertonic-not hypertonic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hypotonic-not hypotonic</td>
<td>.56**</td>
<td>.66**</td>
<td>.57**</td>
<td>-.48**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>unintelligible-intelligible</td>
<td>-.51**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As can be seen in Table 5.8 the perceptual scale slow-quick is not found to be related to any of the acoustic measures, something that could be expected since this scale does not refer directly to voice quality. The scales unsteady-steady and jerking-fluent are related to the standard deviation of the fundamental frequency, a lower standard deviation is related to a more steady and fluent voice.

The median fundamental frequency is related to the perceptual scales concerning pitch (deep-shrill and low-high). Jitter is related to the perceptual scale bubbly-not bubbly only, and not to any other perceptual scales regarding voice quality. The harmonics-to-noise ratio and percentage of voiced both show correlations with the three evaluative scales (deviant-normal, ugly-beautiful and unpleasant-pleasant) and also with some perceptual scales regarding voice quality (bubbly-not bubbly, breathy-not breathy, weak-powerful and monotonous-melodious). The glottal-to-noise excitation ratio seems to be related especially to the perceptual scales involved with the tonicity of the neoglottis (bubbly-not bubbly, breathy-not breathy, weak-powerful, hypertonic-not hypertonic and hypotonic-not hypotonic). The band energy difference is related to all perceptual scales concerning voice quality.

5.3.7 Maximum phonation time versus perceptual evaluations

In section 5.3.7.1 the relations between the maximum phonation time and the overall judgment of voice quality by trained raters as good, reasonable and poor are described. In section 5.3.7.2 the relations between maximum phonation time and the perceptual scale judgments of the trained raters are described.

5.3.7.1 Maximum phonation time versus overall judgment of voice quality

One-way analysis of variance shows that maximum phonation time is related to the judgment of overall voice quality (p=.013). In the poor group (n=12) the mean maximum phonation time was 9.8 s, for the reasonable group (n=14) it was 10.6 s, and for the good group (n=13) it was 17.9 s. The difference between the reasonable and the good group was significant (p=.032), as was the difference between the poor and the good group (p=.022).
5.3.7.2 Maximum phonation time versus perceptual scale judgments

Pearson’s correlation coefficients show that the maximum phonation time is statistically significant related to the scales jerk-fluent (r=.45; p=.004) and intelligible-unintelligible (r=.34; p=.035).

5.3.8 ACOUSTIC SIGNAL TYPING, ACOUSTIC MEASURES AND MAXIMUM PHONATION TIME VERSUS SOCIODEMOGRAPHIC AND CLINICAL FACTORS

In section 5.3.8.1 relations between the acoustic signal types and sociodemographic and clinical factors (sex, age, postoperative follow-up, myotomy, neurectomy, radical neck dissection, and radiotherapy) are described, in section 5.3.8.2 relations between acoustic measures and the sociodemographic and clinical factors, and in section 5.3.8.3 the relations between the maximum phonation time and sociodemographic and clinical factors.

5.3.8.1 Acoustic signal typing versus sociodemographic and clinical factors

Chi-square tests for linear-by-linear association did not reveal any relations between the sociodemographic and clinical factors and the acoustic signal types. Thus, although the acoustic signal types show large variability amongst the speaker group, no relations were found between the characteristics seen in the acoustic signal types and the sociodemographic and clinical factors.

5.3.8.2 Acoustic measures versus sociodemographic and clinical factors

These relations are investigated by means of a t-test for two independent samples. For the acoustic parameter percentage of voiced the non-parametric Mann-Whitney test was used. First, the influence of reconstruction on the acoustic measures was tested. The 30 patients after standard total laryngectomy were compared with the 9 patients who underwent total laryngectomy combined with partial or full pharyngeal reconstruction. The patients with a standard total laryngectomy showed a higher median fundamental frequency (p=.008; 111 Hz versus 65 Hz), a higher percentage of voiced (p=.036; 73% versus 44%), a larger band energy difference (p=.022; -20.78 dB versus -12.84 dB), and a higher harmonics-to-noise ratio (p=.001; 5.2 dB versus 1.3 dB). The larger extent of surgery in the patients who underwent a partial or full pharyngeal reconstruction, leads obviously to less favorable acoustic characteristics of voice quality and fundamental frequency.

Since the differences between the standard and the reconstruction group are very obvious, the remaining clinical factors were studied within the standard group only. Regarding radiotherapy, it was found that the band energy difference tended to be larger in the patient group, receiving radiotherapy as primary treatment of the tumor and laryngectomy for recurrence of the disease, than in the patient group irradiated postoperatively (p=.055; -23.81 dB versus -17.39 dB). Only two patients did not receive radiotherapy at all, these two patients were left out of the analysis for statistical reasons.

For sex, age, postoperative follow-up, myotomy and neurectomy no relations with the acoustic measures were found.

5.3.8.3 Maximum phonation time versus sociodemographic and clinical factors

For maximum phonation time, as for the acoustic measures, first the influence of reconstruction was tested, and then the other clinical factors were tested within the standard group only.

For radiotherapy, a relation was found: the patient group who underwent primary radiotherapy and were laryngectomized for recurrence of the disease, had a longer maximum phonation time than the patient group that received postoperative radiotherapy (p=.009, 16 s
versus 9 s). Only two patients did not receive any radiotherapy at all, and as before, these two patients were left out of the analysis for statistical reasons. This finding is in concordance with the results of the perceptual evaluations (see Chapter 4). Also for radical neck dissection a relation was found, the patients who underwent a radical neck dissection had shorter maximum phonation times than the patients who did not (p=0.049, 9 s versus 14 s).

Reconstruction, sex, age, postoperative follow-up, myotomy and neurectomy were not found to be related to maximum phonation time.

5.4 DISCUSSION

The aim of the present study was to search for valid and reliable objective acoustic parameters that can be obtained for the entire range of tracheoesophageal voice qualities and to investigate relations between perceptual evaluations and acoustic measures (periodicity and spectral analysis) of tracheoesophageal voice quality. Also relations between clinical factors and acoustic measures are studied. In this section the results described in section 5.3 will be discussed in the same order.

5.4.1 ACOUSTIC SIGNAL TYPING

The use of acoustic signal typing is adopted from Titze (1994) and on the basis of the narrow-band spectrograms it was adapted for acoustic signal typing of tracheoesophageal voice. Acoustic signal typing resembles a subjective use of narrow-band spectrograms for classification of tracheoesophageal voice. Despite its subjective character it is an important part of the study since the use of clear criteria and examples provides a relatively easy way of classification. The visual differences between the narrow-band spectrograms are obvious and provide direct insight into the acoustic characteristics of the voice.

5.4.2 ACOUSTIC MEASURES

The choice to use the software program Praat for the acoustic analyses (periodicity and spectral analyses) performed in the present study lies simply in the fact that with this program the fundamental frequency measures can be computed in a more reliable way for a larger part of the speaker group. Additionally, four acoustic measures can be computed for the entire patient group with this program (harmonics-to-noise ratio, percentage of voiced, glottal-to-noise excitation ratio and band energy difference). This means that an objective measure of voice quality can be obtained for each speaker and that even for the poor speakers acoustic measures can be related to the results of perceptual evaluations.

Like in a number of other studies (Debruyne et al., 1994; Moon and Weinberg, 1987; Robbins et al., 1984; Van As et al., 1998b), in the present study the acoustic measures show a large range and standard deviation pointing to a large variability amongst the speakers. The mean fundamental frequency of 104 Hz found in this study, is comparable with the mean fundamental frequency of normal male speakers who are known to have fundamental frequencies around 110 Hz (Van As et al., 1998b). Fundamental frequencies for tracheoesophageal speech found in other studies are 115 Hz by Van As et al. (1998b), 83 Hz by Robbins et al. (Robbins et al., 1984), 92 Hz by Bertino et al. (1996), between 50 Hz and 110 Hz by Debruyne et al. (1994), and between 33 Hz and 121 Hz by Moon and Weinberg (1987). Although it could indeed be said that the mean fundamental frequency of normal male speech (110 Hz) and tracheoesophageal speech is comparable considering male patients (109 Hz on average), for female patients the fundamental frequency is much lower (115 Hz on average) than normal female speech (220 Hz). We agree with Moon and Weinberg (1987) that the high degree of inter-subject variation in fundamental frequency is an important characteristic of tracheoesophageal speech. The high inter-speaker variability is in
concordance with other results that are presented in this thesis; a wide variability is found for the perceptual evaluations, videofluoroscopy results and results of digital high-speed imaging, as well. This is certainly also a result of the formation of the speaker group studied: it was meant to include the entire range of tracheoesophageal voice qualities and even some speakers after partial or full pharyngeal reconstruction were included. Apparently the inter-speaker variability amongst tracheoesophageal speakers is larger than amongst 'normal' speakers, due to a larger anatomical and morphologic variation of the neoglottis compared to the vocal folds. The mean and standard deviation found for the band energy difference is in concordance with the values found by Debruyne et al. (1994).

5.4.3 Maximum Phonation Time

For maximum phonation time, the inter-speaker variation is also high (range 3 to 37 s, average 13 s), these results are comparable with results found in our earlier studies (Van As et al., 1998a; Van As et al., 1998b) and also to the results of Robbins et al. (1984) and Baggs and Pine (1983). Debruyne et al. (1994) report a shorter mean maximum phonation time for tracheoesophageal speech of 6 s. The average maximum phonation time of normal male speakers was found to be 26 s (Van As et al., 1998b), the average of the tracheoesophageal speakers is thus far lower, although some tracheoesophageal speakers can produce maximum phonation times within the normal range.

5.4.4 Acoustic Signal Types versus Acoustic Measures and Maximum Phonation Time

Relations between the acoustic signal types and the acoustic measures show that for only 1 out of the 8 voice samples in the type-IV group fundamental frequency measures can be computed. In this voice sample only a short part of the voice sample contained one harmonic, the percentage of voiced was 18%, indicating that indeed the majority of the voice sample does not contain a periodic voice signal. From this, the rule can be formed that fundamental frequency measures should not be computed in type-IV voice samples. In the type-I group, on the other hand the percentage of voiced was always 100%, separating this type clearly from the other types too. As can be expected, the median fundamental frequency is not found to be related to the acoustic signal types. Surprisingly, also the glottal-to-noise excitation ratio, a measure that is expected to reflect roughness in a voice, is not found to be related to the acoustic signal types. The standard deviation of the fundamental frequency is remarkably higher in the type-III group, and is thus probably reflecting the instability of the fundamental frequency. The percentage of jitter is remarkably lower in the type-I group, indicating that the fundamental frequency is most stable in this group. The harmonics-to-noise ratio and the band energy difference show a clear relationship with the four acoustic signal types.

5.4.5 Acoustic Signal Typing versus Perceptual Evaluations

The four defined types of acoustic signal typing in relation to the overall perceptual judgment of voice quality show that a type-I signal is never perceived as poor and a type-IV signal never as good. The clear relationship between this acoustic signal typing and the perceptual impression of overall voice quality provides the evidence for the use of acoustic signal typing in tracheoesophageal voice. Not only type-I signals are perceived as good, also 50% of the type-II signals is perceived as good. A possible explanation for this might be that the noise masks the deviancy of the voice source signal. Two type-III signals are also perceived as good. Closer inspection of the narrow-band spectrograms of those two speakers shows that they almost met the criterion of a stable harmonic for longer than two seconds. Apparently, for the perception of voice quality in read-aloud text, this criterion is not that important.
Relations between the acoustic signal types and the perceptual semantic bipolar 7-point scales judged by the trained expert raters, give good insight in the relations between the visible acoustic characteristics and specific perceptual scales. Type-IV voice samples are mainly related to the perceptual scales that are related to tonicity (weak-powerful, bubbly-not bubbly, breathy-not breathy, hypotonic-not hypotonic). This is not surprising since the type-IV signals show no or only very short detectable harmonics, which is most probably caused by the inability of a hypotonic neoglottis to produce a periodic voice sound.

Another interesting relation is seen with the scale unintelligible-intelligible; voice samples with less favorable acoustic characteristics show a decreased intelligibility. This suggests a relation between voice quality and intelligibility, something that might be explained by two mechanisms. First, the neoglottis plays a role in the production of the voiced/voiceless distinction: less favorable acoustic characteristics point to a poorer functioning neoglottis and probably thereby poorer ability to produce the voiced/voiceless distinction. Second, the deviant character of the voice quality might negatively influence intelligibility. The relation between intelligibility and voice quality was confirmed in a Master’s thesis by Boon-Kamma (2001), who studied intelligibility of words and spontaneous speech of the same speaker group that participated in the present study, by means of transcription; the results were found to be related to the judgments of overall voice quality (see Chapter 4). Hypertonicity shows a relation with the type-III signals. The unsteady component of the type-III signals that do contain harmonics up to 1000 Hz (and could thus be considered type-I signals, if the instability were not taken into account) may be caused by hypertonicity of the neoglottis leading to a decreased control of voice production. Inspection of videofluoroscopy recordings (see also Chapter 6) of those three speakers shows a hypertonic neoglottis in two of them. In the third speaker the tonicity of the neoglottis based on the videofluoroscopy recordings is judged as hypotonic, closer inspection of the videofluoroscopy recordings showed that although the neoglottis is not hypertonic, the voice sounds strained and hypertonic due to the fact that the esophageal side of the voice prosthesis is pushed into the posterior wall of the esophagus during stoma occlusion for phonation. In the type-I signals the mean rating for tonicity is lower (more hypertonic) than for the type-II and type-IV signals, in the type-III signals it is lowest. This can be explained by the fact that the extreme hypertonic voices fall into group III since they show instability, whereas the voices with slight hypertonicity are considered type I, since the tonicity causes more and clearer harmonics in the spectrogram. Also the scales noise-no noise, jerking-fluent and unsteady-steady show lower values for the type-III signals, which is in concordance with the criterion of the inability to sustain stable harmonics for longer than two seconds. The contribution of the scale noise-no noise also makes sense in this respect, since this scale judges speaking noise caused by incomplete or incorrect timing of stoma occlusion, which then causes instability of the voice.

5.4.6 ACoustIC MEASURES VERSUS PERCEPTUAL EVALUATIONS

Some of the acoustic measures show relations to the overall perceptual judgment as good, reasonable or poor. The median fundamental frequency was not found to be related to overall voice quality. The standard deviation of the fundamental frequency clearly decreases with better voice quality, and the band energy difference, percentage of voiced and harmonic-to-noise ratio show better values for the better voice groups. The percentage jitter was not found to be related to the overall voice quality. This might also be caused by the fact that jitter values for the tracheoesophageal speakers are, on average, relatively high and should therefore be considered unreliable according to Titze and Liang (1993), who found perturbation measures above 5% to be unreliable. As was the case in relation to the acoustic
signal types, the *glottal-to-noise excitation ratio* was not found to be related to the overall judgment of voice quality.

Pearson's correlation coefficients between the perceptual scale judgments and the acoustic measures showed moderate to strong correlations. The scales *noise-no noise, jerking-fluent, and unsteady-steady* are related to the *standard deviation of the fundamental frequency*. This can be explained by the phenomenon seen in the type-III signals: an unstable voice signal, caused by hypertonicity and/or insufficient stoma occlusion, leads to a more fluctuating fundamental frequency and thus a higher standard deviation of the fundamental frequency. The scale *slow-quick* was not found to be related to any of the acoustic measures, something that could be expected since this scale reflects tempo. The perceptual scales related to pitch indeed correlate with the acoustic measures concerning fundamental frequency, something that was not found in our pilot study (Van As et al., 1998b). In this respect, it should be noted that in that pilot study also fundamental frequency results of very short analyzable parts of voice samples that could otherwise be seen as aperiodic were included, thereby creating a more positive and better impression of the fundamental frequency. The *harmonics-to-noise ratio* and the *percentage of voiced* are both related to the same perceptual scales, although for the *percentage of voiced* stronger correlations were observed. The *band energy difference* is related to the same scales as *harmonics-to-noise ratio* and *percentage of voiced* are, but the scales *expressionless-expressive* and *unintelligible-intelligible* show moderate correlations with this parameter too. Correlations with the *band energy difference* are all negative, indicating that the larger the band energy difference, the lower the relative intensity of high-frequency noise and thus the better the judgment of voice quality scales. The *jitter* shows a moderate correlation with the scale *bubbly-not bubbly*: the more bubbly the voice sounds, the higher the jitter value. It should, however, be kept in mind that for the voice samples of the poorest voices no measures based on fundamental frequency could be computed. The correlations for these measures are thus based on the ‘better’ voices.

### 5.4.7 Maximum Phonation Time Versus Perceptual Evaluations

Relations between *maximum phonation time* and the overall judgment of voice quality and those between maximum phonation time and the perceptual scales, indicate that as simple a measure as maximum phonation time is, it is nevertheless an important aspect in tracheoesophageal speech. The better the overall voice quality judgment, the longer the maximum phonation time. Short maximum phonation times are related to less fluent and less intelligible tracheoesophageal speech. In this respect it might be the case that in tracheoesophageal speech, maximum phonation time is more important than in normal speech. The mean maximum phonation time is considerably shorter for tracheoesophageal speech than for normal speech (Van As et al., 1998b), although there is some overlap between both speaker groups. Probably the tracheoesophageal speakers that have maximum phonation times close to that of normal speakers are perceived as better. Another explanation for this might be that speakers with a better voice quality and consequently a better functioning neoglottis are able to produce longer maximum phonation time due to a better closure of the neoglottis.

### 5.4.8 Acoustic Analyses Versus Sociodemographic and Clinical Factors

As for the perceptual evaluations, the obvious better voice quality for the patient group after standard total laryngectomy compared to the patient group that underwent a partial or full pharyngeal reconstruction, is also reflected in the acoustic measures. The patient group after standard total laryngectomy had higher *fundamental frequencies*, higher *percentages of voiced*, a larger *band energy difference*, and a larger *harmonics-to-noise ratio*. The *glottal-to-
Noise excitation ratio appeared not to be suitable for distinguishing these two speaker groups. An interesting influence was found for radiotherapy: the maximum phonation time was longer and the band energy difference was larger in the patient group that received primary radiotherapy and underwent a total laryngectomy for recurrence of the disease. An explanation for this might be that in the patient group that received postoperative radiotherapy the field of radiation was larger and there is usually more tissue fibrosis in his group. Also for radical neck dissection an influence was found: the maximum phonation time was longer in the patient group that needed no neck dissection. A possible explanation for this might be that patients who had a radical neck dissection show more oedema of the tissues and thereby probably stiffer esophageal mucosa. One must however be careful in drawing conclusions from this: both radiotherapy and radical neck dissection are important factors for optimal tumor control. More precise investigations, taking the exact radiation fields and precise surgical procedures of radical neck dissection into account, might provide more precise information on these relations. Regarding sex the absence of a difference between the fundamental frequency of male (109 Hz on average) and female (115 Hz on average) tracheoesophageal speech is confirmed. In normal voices an evident difference between male and female voice would be recognized. Most studies of tracheoesophageal speech consider male patients only. The reason for this is mainly a practical one: more males develop laryngeal cancer. Trudeau and Qi (1990) studied 10 female esophageal speakers and concluded that characteristics of tracheoesophageal speech may be highly similar regardless of speaker gender. The psychosocial implications of such a male sounding voice for a female speaker are self-evident.

Regarding the remaining clinical factors age, postoperative follow-up, myotomy, and neurectomy, no significant relations with the acoustic measures were found.

5.5 Conclusions

In this Chapter the usefulness of acoustic signal typing of tracheoesophageal voice has been established. A narrow-band spectrogram gives useful information on the acoustic characteristics of the voice sample and the four different signal types show clear relations with the overall perceptual judgment, with the majority of the perceptual scale judgments regarding voice quality and with the acoustic measures regarding voice quality.

Using Praat, the fundamental frequency and other acoustic measures based on fundamental frequency (like perturbation measures) could be computed for 77% of the patients in the group studied. The remaining 23% of the voice samples show no periodicity at all and consequently no measures of fundamental frequency can be obtained for these patients. Four acoustic measures could be computed for the entire patient group. From these four measures, especially the band energy difference, the harmonic-to-noise ratio and the percentage of voiced appear to be of interest. Those measures show clear relations with the overall judgment of voice quality, as well as with the majority of the perceptual scales regarding voice quality. Pearson’s correlations found with the perceptual scales have shown to be moderate to strong. The band energy difference appeared also to be useful in relation to the clinical factors of reconstruction and radiotherapy. The glottal-to-noise excitation ratio appeared to be of no interest for evaluation of overall voice quality and evaluative perceptual voice quality scales. Pearson’s correlation coefficient showed however, that this acoustic parameter seems to be related especially to the perceptual scales regarding tonicity, and could thus be of interest in that respect. Furthermore, the use of maximum phonation time as an objective measure has shown to be relevant in relation to tracheoesophageal voice quality.
CONCLUDING REMARKS

In the present Chapter acoustic analyses of tracheoesophageal speech have been performed and related to the results of perceptual evaluations. It can be concluded that with Praat fundamental frequency can be extracted correctly when periodicity is present in the voice sample according to the narrow-band spectrogram. Acoustic signal typing is very informative for tracheoesophageal voice quality. Correlations between the perceptual evaluation of pitch and the measurement of fundamental frequency are strong and can thus be considered useful as an objective measure of pitch. The acoustic measures of percentage of voiced, harmonics-to-noise ratio and band energy difference show moderate to strong correlations with the perceptual scales regarding voice quality and can thus be considered useful as an objective diagnostic tool of voice quality. The glottal-to-noise excitation ratio has been shown to be a useful objective parameter in relation to the perceptual judgment of tonicity. Also, maximum phonation time has shown to be a relevant measure in relation to tracheoesophageal voice quality judgment.

In the following chapters, the perceptual evaluations and acoustic measures of voice quality will be completed with investigations of the anatomical and morphologic characteristics of the neoglottis. In Chapter 6 investigations of neoglottic characteristics by means of videofluoroscopy are described, followed by Chapter 7 in which the neoglottic characteristics studied by videofluoroscopy are related to the perceptual evaluations of voice quality (described in Chapter 4) and to the results of acoustic analyses described in the present chapter. In Chapter 8 investigations of neoglottic characteristics by means of digital high-speed imaging are described, followed by Chapter 9, in which the neoglottic characteristics studied by digital high-speed imaging are related to the perceptual evaluations (described in Chapter 4), to the acoustic analyses described in the present chapter, and to the neoglottic characteristics studied by videofluoroscopy (Chapter 7).

REFERENCES


